·基础理论与方法·

判定传染病发病时间聚集性五种方法的比较与探讨

孙建伟　许汴利　陈豪敏　康锴　黄丽莉

【导读】 探讨并比较判定传染病发病时间聚集性的五种方法。以郑州市金水区2008—2010年细菌性痢疾发病时间聚集性的判定为例，比较可用于传染病发病时间聚集性分析和判定的聚类分析、游程检验、负二项分布、圆分布、集中度五种方法。聚类分析得出5—9月是发病流行期，8月是发病高峰期；游程检验得出2008和2009年月发病具有聚集性（$P<0.05$），2010年发病月分布具有随机性；集中度法得出2008—2010年每年月发病分布均有一定的聚集性，从M值看聚集程度逐年减弱；圆分布法得出2008—2010年每年均具有发病聚集倾向时间（$P<0.01$），3年依次为7月11、29日和8月24日；负二项分布拟合得出在年度日发病层面2008和2010年具有聚集性（$P>0.05$），2009年不具有聚集性（$P<0.05$）。五种方法均可用于传染病发病时间聚集性的分析和判定，可根据不同情况选择。

【关键词】 传染病；时间聚集性；聚类分析；游程检验；负二项分布；圆分布；集中度

Discussion on five methods used for the determination of temporal clustering on infectious diseases

SUN Jian-wei, XU Bian-li, CHEN Hao-min, KANG Kai, HUANG Li-li. Institute of Infectious Disease, Henan Provincial Center for Disease Control and Prevention, Zhengzhou 450016, China
Corresponding author: XU Bian-li, Email: bianlixu@yahoo.cn

【Introduction】 To demonstrate and evaluate five different methods in the determination of temporal clustering on infectious diseases. The incidence rates of bacillary dysentery in Jinshui district, Zhengzhou city, Henan province from 2008 to 2010 were analyzed by 5 different methods—Cluster Analysis, Runs Test, Negative Binomial Distribution, Circular Distribution and Concentration Ratio. Through Cluster Analysis, data showed that the epidemic period was from May to Sept. with August as the peak. Runs Test confirmed a cluster of month-incidence in 2008 and 2009 ($P<0.05$) and a random distribution in 2010. The Concentration Ratio showed a weakened seasonal incidence cluster to a certain extent by M from 2008 to 2010. The Circular Distribution demonstrated an inclining cluster of time ($P<0.01$) and it was on July 11th and 29th, as well on August 24th in 2008, 2009 and 2010. In terms of day-incidence, the Negative Binomial Distribution presented a cluster in 2008 and 2010, but with no significant difference in 2009. The five above said methods could flexibly be used in determining the temporal clustering of infectious disease at different occasions.

【Key words】 Infectious disease; Temporal clustering; Cluster analysis; Runs test; Negative binomial distribution; Circular distribution; Concentration ratio

传染病发病时间聚集性的判定是流行病学研究的重要内容之一，既可为病因研究提供线索，又可为暴露时间或潜伏期研究提供支持，还可用于评价不同时期传染病防控效果[1]，为制定有效防控措施和卫生决策提供科学依据。在发病时间聚集性判定和分析方面，国内目前常用的方法有图示法、集中度、圆分布法[2,3]，而聚类分析、负二项分布、游程检验则应用较少，且未见有方法间的比较和讨论。本研究以郑州市金水区2008—2010年细菌性痢疾（菌痢）发病时间数据为例，除集中度、圆分布法外，又试用聚类分析、游程检验、负二项分布对发病时间数据进行分析，并比较和探讨五种方法各自在传染病发病时间聚集性判定中的适用范围和价值。

基本原理

1. 聚类分析[4]：是统计学研究事物分类的一种方法，在关注对象分类面貌尚不清楚情况下，通过彼此之间比较而将性质相近的归为一类，性质差别较大的归为不同类。在Q型聚类（样品聚类）中衡量对象性质相近程度的距离指标常用欧式距离，其计算

DOI: 10.3760/cma.j.issn.0254-6450.2011.10.017
作者单位：450016 郑州，河南省疾病预防控制中心传染病预防控制所
通信作者：许汴利，Email: bianlixu@yahoo.cn

公式：

$$d_{ij}=\sqrt{\sum_{k=1}^{m}(X_{ik}-X_{jk})^2}$$

式中X_{ik}表示第i个对象的第k个指标的观察值，X_{jk}表示第j个对象的第k个指标观察值；m是观察指标的总数。d_{ij}表示第i个对象与第j个对象间的距离，d_{ij}越小，对象"性质"越接近。在实例分析中，对象为月份，指标是不同年份，观察值是发病数。

2. 游程检验[5]：亦称"连贯检验"，是根据样本标志表现排列所形成的游程的多少进行判断的检验方法，可用于检验样本的随机性。如某样本$n=12$的标志表现为男、女，有3种排列：①男\男，女\女\女，男，女\女，男\男\男\男；②男\男\男\男\男\男\男，女\女\女\女\女；③男，女，男，女，男，女，男，女，男，女，男\男。连续出现男或女的区段称为游程，每个游程包含的个数为游程长度。以r表示序列中游程的个数，其3种排列分别为：$r=5$、$r=2$、$r=11$。可见①是随机性序列；②、③是非随机性序列。所以可用游程的个数(number of runs)来检验样本的随机性，用来观察某一传染病的发生是随机分布还是具有聚集性。如在某一关注的时间范围内对某种传染病发病按时间顺序以不同时间单位(月、周、日等)发病数为序列，以中位数(M)或平均数(或其他分界值)将发病数据分为$\geq M$和$<M$的两种情况，用上下交错形成的游程个数来检验样本是否随机。

3. 负二项分布[6]：在医学中主要用于聚集性疾病及微生物、寄生虫分布模型等的研究，有2个参数即μ和k。参数μ一般用样本均数$\bar{x}$作为其估计值，聚集性参数k值本研究用矩法$k=\bar{x}^2/s^2-\bar{x}$求得$\hat{k}$。然后根据$p=\mu/k$，$q=1+p$计算出p和q值，再根据

$$\begin{cases}p(0)=q^{-k} & X=0\\ p(X)=\dfrac{(k+x-1)p}{xq}P(x-1) & X\geq 1\end{cases}$$

计算出理论概率$p(X)$。通过$T=NP(X)$计算出理论频数、$\sum\dfrac{(T-f)^2}{T}$计算出拟合检验的χ^2值，通过查χ^2界值表确定P值范围。

4. 集中度[7]：表示发病时间季节性强弱的指标，由各月发病数与全年发病总数之比通过下式计算：

$$\begin{cases}R_x=(r_2+r_6-r_8-r_{12})/2+\sqrt{3}(r_3+r_5-r_9-r_{11})/2+(r_4-r_{10})\\ R_y=(r_3-r_5-r_9+r_{11})/2+\sqrt{3}(r_2-r_6-r_8+r_{12})/2+(r_1-r_7)\end{cases}$$

$$M=\sqrt{R_x^2+R_y^2}$$

M表示集中度，R表示离散度，r_i表示月发病数与全年发病总数之比。$M=1$时说明病例集中在1个月内；在0.9以上说明发病有严格季节性；在0.7～0.9之间，说明发病有很强的季节性；在0.5～0.7之间说明发病有较强季节性；在0.3～0.5之间说明发病有一定季节性；在0.3以下说明发病时间分布比较均匀；$M=0$表示病例均匀分布在12个月内。

5. 圆分布法[7,8]：是将具有周期性变化的资料通过三角函数变换，使原始数据成线性资料的一种统计学方法。一组圆分布资料如果有集中分布的倾向，这一倾向性可用平均角表示。将1年(365 d)转化为360°，1 d则对应0.9863°，以每月中间一天为组中值，换算成对应的角度，如1月16日对应$16\times0.9863°=15.78°$，2月14日对应$(31+14)\times0.9863°=44.38°$，3月16日对应$(31+28+16)\times0.9863°=73.97°$，余类推。通过以下公式求得平均角$\bar{\alpha}$，换算成对应的月日即为发病高峰期时间。

$$\begin{cases}X=\sum f\cos\alpha/n\\ Y=\sum f\sin\alpha/n\\ r=\sqrt{X^2+Y^2}\end{cases}$$

$$\cos\bar{\alpha}=X/r \qquad \sin\bar{\alpha}=Y/r$$

平均角$\bar{\alpha}$的检验是计算雷氏Z值(Rayleigh Z)：$Z=nr^2$，通过雷氏Z值临界值表判定平均角$\bar{\alpha}$有无统计意义。平均角$\bar{\alpha}$的标准差$s=\dfrac{180}{\pi}\sqrt{-2\ln r}$，可用于发病时间可信区间的估算。

6. 统计学分析：聚类分析、游程检验通过SPSS 17.0软件完成[5]，负二项分布、集中度、圆分布通过相应公式计算得出。

实例分析

以郑州市金水区2008—2010年菌痢分月与分日发病数据分析为例，进行发病时间聚集性的判定和分析。该区3年菌痢发病数分别为409、219和194例，5—9月的流行期发病数所占比例分别为71.64%、69.41%和63.92%，发病数最多月份3年分别为7、8和8月(图1)。

1. 聚类分析：执行SPSS"分析 | 分类 | 系统聚类"，以各年度每月发病数为"变量"，选择"作图"中的"系统树图"(图2)。若聚为两类，则5—9月为一类，1—4月及10—12月为一类。高发期5—9月可再分为两类，8月为一类，5—7月及9月为一类；若分为三类，则5月和9月为一类，6—7月为一类，8月为一类。

2. 游程检验：执行SPSS"分析 | 非参数检验 | 游程"，以各年度每月发病数为"检验变量列"，以平均数(即月平均发病例数)为"检验值(test value)"进

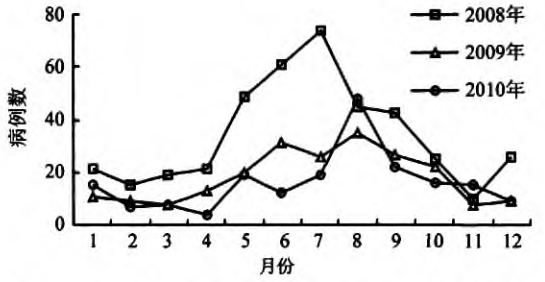

图1　2008—2010年郑州市金水区菌痢病例月分布

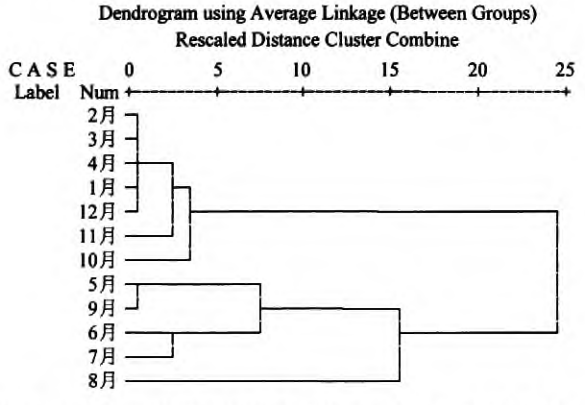

图2　2008—2010年郑州市金水区菌痢发病分月聚类分析

行各年度月发病数序列的游程检验，结果见表1。如2010年月发病分布以“+”代表≥月平均发病例数、“—”代表<月平均发病例数，则年度月发病序列为“—\—\—\—，+，—，+\+\+，—\—\—”，游程数为5，结果$P>0.05$，即若以月平均发病数为检测值则可认为2010年菌痢月发病符合随机分布，即尚不能认为具有聚集性。

3. 负二项分布拟合：根据全年每日发病数据，按日发病例数0～7例分类统计计数。结果见表2。2008和2010年每日发病数服从负二项分布（$P>0.05$），具有时间聚集性，而尚不能认为2009年日发病具有聚集性（$P<0.05$）。

4. 圆分布法与集中度法：结果见表3。由集中度结果可见2008—2010年月发病均为有一定的聚集性，且聚集程度逐年下降；由圆分布法结果可认为每年度均具有发病时间聚集性（$P<0.01$），且3年来发病集中时间及流行期有逐年后移现象。

表1　2008—2010年郑州市金水区菌痢发病月分布游程检验

年份	检验值	病例数 ≥检验值	病例数 <检验值	游程个数	Z值	P值
2008	34.08	5	7	3	−2.082	0.037
2009	18.25	6	6	3	−2.119	0.034
2010	16.17	4	8	5	−0.575	0.565

讨　论

经文献检索，目前未见有传染病发病时间聚集性的一般定义与判定标准。传染病发病时间聚集性判定至少应考虑三个方面。一是所关注的时间范围，例如观察的是连续多年，还是某年、某月，甚或是可能跨年度的某一流行期等；二是所关注的时间层面或时间单位，如同是某一年发病时间数据，可从月、旬、周、日等不同时间层面进行分析；三是判定所用方法，同一组数据用不同分析方法所得出结论不尽相同，表现为聚集性的有或无以及聚集程度的大小。故在作出判定结论时，首先应是统计结果与该传染病流行病学的有机结合，其次是结论中应包括以上三个要素。

由于以上三个不同因素所致，在2008—2010年郑州市金水区菌痢发病时间聚集性判定实例分析中，五种方法从不同角度得出不同结论，且结论侧重点各不相同。聚类分析侧重于按发病数对3年来各月份进行不同层次的归类，游程检验对各年度月发病数的排列是否具有随机性进行检验，负二项分布则以年度每日发病数归类统计进行拟合，圆分布法则侧重于年度发病聚集倾向时间即$\bar{\alpha}$所在时间的判定，集中度则可将聚集性通过M值进行量化。从方法的繁易程度看，聚类分析和游程检验可直接通过

表2　2008—2010年郑州市金水区菌痢发病日分布负二项分布拟合

发病数（X）	2008年 实际日数（f）	2008年 理论日数（T）	2008年 $(T-f)^2/T$	2009年 实际日数（f）	2009年 理论日数（T）	2009年 $(T-f)^2/T$	2010年 实际日数（f）	2010年 理论日数（T）	2010年 $(T-f)^2/T$
0	155	148.04	0.33	214	208.05	0.17	228	229.18	0.01
1	98	109.17	1.14	95	109.35	1.88	97	93.91	0.10
2	57	58.91	0.06	45	35.55	2.51	28	29.99	0.13
3	35	27.89	1.81	10	11 2.05	0.09	8	12 11.91	0.00
4	11	12.29	0.14	1			3		
5	7	9 8.69	0.01				1		
6	1								
7	1								
χ^2值			3.49			4.65			0.24
df			3.00			1.00			1.00
P值			0.32			0.03			0.62
$\hat{k}$			2.16			4.19			1.79

表3 2008—2010年郑州市金水区菌痢发病时间的圆分布法与集中度法拟合

年份	圆分布法				集中度法	
	$\bar{\alpha}$所在日期	Z值	P值	68.27%CI	M	月发病聚集性
2008	7月11日	47.70	<0.01	4月17日至10月4日	0.46	有一定聚集性
2009	7月29日	24.30	<0.01	5月4日至10月23日	0.41	有一定聚集性
2010	8月24日	21.04	<0.01	5月29日至11月19日	0.40	有一定聚集性

SPSS、SAS等快速实现，最为简便；其次是集中度分析；而圆分布法和负二项分布的计算或统计软件的编程相对繁琐。从适用的时间范围看，集中度分析最为局限，其余4种方法适用时间范围则广泛、灵活。从方法优点看，聚类分析更适宜于某一疾病连续多年疫情数据时间分布特征的分析，可得出所关注时间单位的不同聚类结果[9,10]；游程检验可用于任何观察时间单位发病数排列是否随机以进行聚集性判定[11]，比负二项分布简便易行，但需要注意分界值的选择；负二项分布可在时、日、周、年等所关注的不同时间单位层面判定发病是否具有聚集性[12]，若有两组以上结果，还可以通过聚集参数k的大小比较聚集程度，需要注意的是聚集参数k估算以最大似然法最为精确[13]；集中度的最大优点是能区分发病季节性强弱，便于同一疾病不同时期或多种疾病聚集性的变化或比较[14]，如能将关注时间段或其他对象十二等分，则可以扩大集中度法应用范围[15]；圆分布法适宜的时间范围更广、更灵活，尤其是可用于跨年度发病数据分析，关键是所关注的时间范围内单位时间与角度的互算，同时需要进行平均角的雷氏检验，这种方法最大的优点是能推算出发病聚集倾向时间及不同范围可信区间所在，可用于不同时期、不同地区某一传染病发病时间高峰的动态观察和比较。

综上所述，在传染病发病时间聚集性分析和判定中，根据病种的不同、关注时间范围的不同、分析研究目的不同，可灵活采取不同方法或多种方法并用。若是以月份为单位进行年度分析，对于流行性乙型脑炎、菌痢等具有显著发病季节高峰的传染病，用常规的图示法即可，聚类分析、游程分析亦可；对于发病无显著高峰期，首选集中度法判定发病季节性的强弱。若研究关注的是不同年代或不同地区某一传染病发病高峰期的比较及动态变化，则首选圆分布法；若想要比较前后不同时期发病聚集性强弱则可由负二项分布、集中度实现。同时，需要强调的是应以动态、灵活、变化的观点综合看待传染病发病时间聚集性，其分析和判定的主要目的是为传染病发生和流行的相关影响因素研究提供线索，最终服务于传染病防控措施的科学制定与完善。

参 考 文 献

[1] Lv DB, Wang TP, Wu WD, et al. Evaluation on the elimination effect of oncomelania hupensis gredler by negative binomial distribution. J Trop Dis Parasitol, 2003, 1(2): 106-108. (in Chinese)
吕大兵，汪天平，吴维铎，等. 应用负二项分布评价灭螺效果. 热带病与寄生虫学，2003，1(2)：106-108.

[2] Shan RQ, Xu Y, Xue DY. The application of the concentration ratio and the circular distribution in the analysis of seasonality of infectious diseases. Dis Surveil, 2006, 21(11): 589-591. (in Chinese)
山若青，徐毅，薛大燕. 应用集中度和圆形分布分析传染病的季节性. 疾病监测，2006，21(11)：589-591.

[3] Tan RM. The application and comparison of the concentration ratio and the circular distribution in the analysis of temporal clustering of infectious diseases. Mod Prev Med, 2007, 34(1): 74-75. (in Chinese)
谈荣梅. 集中度和圆形分布法在传染病发病时间聚集性分析中的应用比较. 现代预防医学，2007，34(1)：74-75.

[4] Jin PH. Medical Statistics. 2nd ed. Shanghai: Fudan University Press, 2008. (in Chinese)
金丕焕. 医用统计方法. 2版. 上海：复旦大学出版社，2008.

[5] Shi SH, Shi XZ, Li YY. The application of computer on the analysis of medical data. Lanzhou: Lanzhou University Press, 2005. (in Chinese)
时松和，施学忠，李颖琰. 计算机在医学数据分析中的应用. 兰州：兰州大学出版社，2005.

[6] Han XH. Properties and characters of negative binomial distribution and its application in epidemiology. J Mat Med, 2009, 22(2): 138-140. (in Chinese)
韩新焕. 负二项分布的性质特征及在流行病学研究中的应用. 数理医药学杂志，2009，22(2)：138-140.

[7] Gan YB, Liao Z, Cai J. The analysis of seasonality of epidemic cerebrospinal meningitis in Nanchang from 1985 to 2008 by the concentration ratio and the circular distribution. Chin J Stat, 2010, 27(4): 379-380. (in Chinese)
甘仰本，廖征，蔡军. 应用集中度和圆形分布法分析南昌市1985—2008年流行性脑脊髓膜炎发病季节性. 中国卫生统计，2010，27(4)：379-380.

[8] Wang J, Ye DQ. The circular distribution analysis and application. Chin J Dis Control Prev, 2004, 8(2): 160-161. (in Chinese)
王静，叶冬青. 圆分布分析方法及实例应用. 疾病控制杂志，2004，8(2)：160-161.

[9] Li L, Liang Q, Qi X, et al. Cluster analysis on epidemiologic characteristics of hand-mouth-foot disease in Jiangsu province in 2009. Jiangsu J Prev Med, 2010, 21(6): 4-7. (in Chinese)
李亮，梁祁，祁贤，等. 江苏省2009年手足口病流行特征的聚类分析. 江苏预防医学，2010，21(6)：4-7.

[10] Chu HZ, Wu YP, Zhang J. The cluster analysis of the changes in Qingdao's spectrum of illness in the past 17 years. Chin Hospital Stat, 2006, 13(4): 317-319. (in Chinese)
初慧中，邬贻萍，张进. 青岛市17年疾病谱变化聚类分析. 中国医院统计，2006，13(4)：317-319.

[11] Huang BD. The application of runs test in the determination of epidemic tendency of paratyphoid. Thesis Compilation of the China Health Statistics Academic Forum, 2007. (in Chinese)
黄宝定. 游程检验在副伤寒流行趋势分析判断中的应用. 2007年中国卫生统计学术大会论文集.

[12] Li XG. The study on the spatio-temporal distribution of hemorrhagic fever with renal syndrome in Junan county. Chin J Zoonoses, 2009, 25(7): 705-706. (in Chinese)
李学刚. 莒南县肾综合征出血热时间和空间分布研究. 中国人兽共患病学报，2009，25(7)：705-706.

[13] Zha M. Four methods for the calculation of parameter in K negative binomial distribution. J Trop Dis Parasitol, 2000, 29(2): 107-109. (in Chinese)
查明. 负二项分布参数K的四种估算方法. 热带病与寄生虫学，2000，29(2)：107-109.

[14] Liu TX. The analysis of seasonality of legal report infectious diseases by concentration ratio. Chin Primary Health Care, 2004, 18(12): 61-62. (in Chinese)
柳太祥. 法定传染病季节性分布的集中度分析. 中国初级卫生保健，2004，18(12)：61-62.

[15] Yu LG. The application of concentration ratio analysis on the distribution of infectious disease case's age. Chin Hospital Stat, 2001, 8(4): 230-231. (in Chinese)
余录根. 集中度在分析传染病年龄分布中的应用. 中国医院统计，2001，8(4)：230-231.

（收稿日期：2011-04-25）
（本文编辑：张林东）