·基础理论与方法·

求和自回归滑动平均模型结合圆分布法分析脑卒中死亡率动态规律

王德征　江国虹　宋桂德　吴彤宇　潘怡　张颖　张辉

【导读】 通过1999年1月至2006年12月天津市脑卒中逐月死亡率数据，应用圆分布法探讨脑卒中死亡率的季节分布，动态变化规律，建立监测与预测的时间序列模型。通过模型辨识、参数估计及其检验、白噪声检验、模型的拟合度分析等过程，建立求和自回归滑动平均模型(ARIMA)的季节乘积模型$(p,d,q)(P,D,Q)_s$。脑卒中死亡率以年为周期，一年中1月为高发月份。建立ARIMA$(0,1,0)\times(0,1,1)_{12}$模型：$(1-B)(1-B^{12})\ln x_t=0.001+(1-0.537B^{12})\varepsilon_t$。结论：ARIMA乘积模型结合圆分布法是对脑卒中死亡率进行时间序列分析的重要方法；应用该方法可对脑卒中流行趋势及死亡率进行预测，为卫生资源合理分配、公共卫生政策计划制定和防治结果考核提供科学依据。

【关键词】 脑卒中；死亡率；时间序列；求和自回归滑动平均模型模型；圆分布

Autoregressive integrated moving average model and circle distribution analysis of stroke mortality in Tianjin

WANG De-zheng, JIANG Guo-hong, SONG Gui-de, WU Tong-yu, PAN Yi, ZHANG Ying, ZHANG Hui. Tianjin Center for Disease Control and Prevention, Tianjin 300011, China

【Introduction】 To develop a model for forecasting the mortality of stroke in Tianjin, China. The time series of stroke mortality from 1999 Jan. to 2006 Dec. in Tianjin city were subjected. Circle distribution analysis was used to verify the trend of time concentration. Multiple seasonal autoregressive integrated moving average model [ARIMA $(p, d, q)(P, D, Q)_s$], based on model identification, estimation and verification of parameter, and analysis of the fitting of model, was established. Most of the deaths from stroke occurred in January and had a cycle of 12 months. An ARIMA model $(0,1,0)\times(0,1,1)_{12}$ was established $(1-B)(1-B^{12})\ln x_t=0.001+(1-0.537B^{12})\varepsilon_t$. Conclusion: ARIMA & Circle Distribution analysis is an important tool for stroke mortality analysis. Potentially it has a high practical value on the surveillance, forecasting and prevention of stroke mortality.

【Key words】 Stroke; Mortality; Time Series; Auto regressive integrated moving average model; Circle distribution

脑卒中具有较高的死亡率，已经成为一种严重威胁人类生命和健康的疾病。准确系统分析脑卒中死亡率及其变动规律，对正确制定疾病控制规划，合理配置卫生资源，科学考核防治措施具有重要意义。本研究采用求和自回归滑动平均模型(auto regressive integrated moving average, ARIMA)结合圆分布法对1999－2005年天津市脑卒中死亡率动态分布规律进行探讨。

基本原理

1. 圆分布法[1]：一年时间以圆周表示。12个月按照360°划分，每个月$\alpha_i=30°$。$x=(\sum f_i\cos\alpha_i)/\sum f_i$，$y=(\sum f_i\sin\alpha_i)/\sum f_i$，$r=\sqrt{x^2+y^2}$，$S=180\sqrt{2(1-r)}/\pi$，$Z=nr^2$。根据以上公式计算：$x$、$y$值；角度离散度指标$r$值；平均角$\bar{\alpha}$；角度标准差($s$)；推算出平均角$\bar{\alpha}$所在的日期；计算Rayleighps Z值，检验平均角$\bar{\alpha}$有无统计学意义。

2. ARIMA法[2]：采用季节乘积模型，ARIMA $(p,d,q)(P,D,Q)_s$：$\phi_p(B)\Phi_P(B^S)\nabla^d\nabla^D x_t=\theta_q(B)\Theta_Q(B^S)\varepsilon_t$。

(1) 辨识：对原始数据进行分析，通过对序列的观察及数据的自相关系数(ACF)和偏相关系数(PACF)来判断原始数据确定的时间序列是否平稳。如果不平稳，使用对数变换、一般差分(d)与季节差分$(D)_s$的方法以达到平稳。同时确定差分阶数d、D。转化后的序列符合ARMA(p,q)〔ARMA(P,Q)〕模型，然后利用Akaike准则(AIC)和Schwarz准则(SBC)确定自回归阶数p、P，移动平均阶数q、Q。通过条件最小二乘法确定ARMA(p,d,

DOI:10.3760/cma.j.issn.0254-6450.2009.01.023
作者单位：300011 天津市疾病预防控制中心

q)及ARMA(P,D,Q)$_s$序列。

(2) 参数估计和模型检验:检测所估计模型理想的残差序列应为一随机序列,否则,需进一步进行残差分析并加以改进。

(3) 预测:利用所拟合的模型进行预测。

3. 统计学分析:统计软件为SPSS 11.5。

实例分析

1.资料来源及分析目的:采用1999年1月1日至2006年12月31日天津市疾病预防控制中心收集的天津市居民全死因监测数据,应用ARIMA模型结合圆分布法分析脑卒中死亡率动态规律。脑卒中诊断分类按照"疾病和有关健康问题的国际统计分类"(ICD)[3,4]:ICD-9:430,431,432,434,438;ICD-10:I60-I64(除外I63.6),I69,I69.0,I69.1,I69.3,I69.4,I69.8。以实际死亡月份为分析时间点(1999年1月至2005年12月共84个时间点用于建立模型,2005年1月至2006年12月共12个时间点用于模型外推验证)。以2000年世界标准人口计算的各时间点世界标化月死亡率作为分析时间序列。

2.分析结果:由表1可见1999—2005年天津市脑卒中死亡率有下降趋势。

表1 1999—2005年天津市脑卒中死亡粗率及世界标化死亡率(/10万)

监测年	死亡粗率(95%*CI*)	世界标化死亡率(95%*CI*)
1999	156.572(154.000~159.145)	150.632(148.109~153.156)
2000	160.150(157.553~162.746)	149.076(146.571~151.582)
2001	156.388(153.824~158.951)	140.056(137.630~142.482)
2002	136.670(134.278~139.062)	116.744(114.533~118.955)
2003	141.438(139.012~143.863)	115.584(113.391~117.776)
2004	137.828(135.442~140.213)	108.430(106.314~110.546)
2005	141.560(139.151~143.969)	107.145(105.049~109.241)

应用圆分布法分析显示各年度脑卒中死亡高峰多发生于冬季的1月份(1月3—18日),具有明显的集中趋势($P<0.001$),见表2。根据以上分析结果可见,各时间点脑卒中死亡率为非平稳序列,拟对原始数据(标化月死亡率x_t)对数变换(log)后进行一次一般差分($d=1$),然后一次季节差分($D=1, s=12$)使之形成平稳序列。差分后序列趋于平稳,对序列进行白噪声检验$\chi^2=27.33$,$P<0.007$,表明该序列是平稳非白噪声序列,说明还蕴藏着相关信息,需要提取出来,因此该序列有一个ARIMA模型。

由ACF和PACF图(图略),可见在lag为1和12附近有两个大的峰,lag为1附近ACF、PACF均截尾,lag为12附近ACF截尾PACF拖尾,因此初步确定模型参数$Q_{12}=1$即形式可定为ARIMA(0,1,0)×$(0,1,1)_{12}$。

采用条件最小二乘法对参数进行估计,结果显示MA(1,1)有统计学意义,拟合优度统计量AIC和SBC都较小。即模型ARIMA(0,1,0)×$(0,1,1)_{12}$拟合得较好(表3)。

表2 1999—2005年天津市脑卒中死亡圆分布计算表

年份	频数	r	s	$\bar{\alpha}$	$\bar{\alpha}$对应日期	Z值	P值
1999	14 211	0.079	77.781	40.734	01-11	87.692	<0.001
2000	14 591	0.082	77.651	32.490	01-03	97.200	<0.001
2001	14 278	0.106	76.602	40.455	01-11	161.267	<0.001
2002	12 526	0.107	76.584	47.771	01-18	142.570	<0.001
2003	13 048	0.093	77.157	33.187	01-03	113.548	<0.001
2004	12 808	0.119	76.052	39.926	01-10	181.582	<0.001
2005	13 249	0.097	76.987	33.293	01-03	125.343	<0.001

表3 ARIMA模型的条件最小二乘法参数估计

ARIMA模型	参数	估计值	$s_{\bar{x}}$	t值	P值	Lag	AIC	SBC
$(0,1,0)\times(0,1,1)_{12}$	MU	0.002	0.007	0.23	0.820	0	−125.325	−120.8
	MA1,1(q)	0.537	0.106	5.06	0.000	12		

模型拟合后对残差为白噪声原假设进行检验,各个延迟期χ^2值均较大(lag=12,$\chi^2=17.64$,$P=0.090$;lag=24,$\chi^2=24.9$,$P=0.356$)。所以不能拒绝拟合模型的残差为白噪声,说明残差中蕴涵的信息已经完全被提取出来了。模型为:ARIMA(0,1,0)×$(0,1,1)_{12}$,$(1-B)(1-B^{12})\ln x_t=0.001+(1-0.537B^{12})\varepsilon_t$,其中,$x_t$为月死亡率(世界标化),$B$为后移算子,$\varepsilon_t$为随机干扰。

由ARIMA(0,1,0)×$(0,1,1)_{12}$模型对1999—2005年间逐月脑卒中死亡率时间序列数据的拟合情况以及对2006年逐月死亡率的外推预测情况(图1、表4),可以看出模型预测值的动态趋势与实际情况基本一致,各月份的预测值都落入了预测值的可信区间范围,同时高发月份为1月,预测结果令人满意。

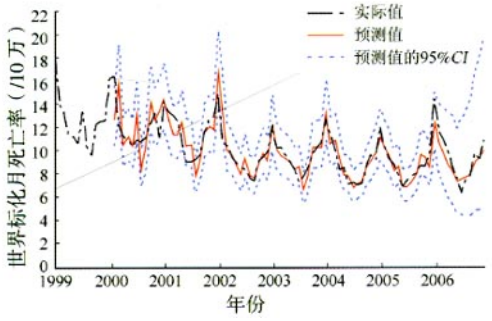

图1 1999—2006年天津市脑卒中世界标化月死亡率实际值、预测值的动态趋势

表4　应用模型预测2006年天津市脑卒中世界标化月死亡率(/10万)

月份	实际值（95%*CI*）	预测值（95%*CI*）	残差
1	14.58(13.74～15.42)	12.50(10.30～15.15)	2.08
2	11.86(11.08～12.63)	10.70(8.13～14.10)	1.16
3	11.05(11.30～11.81)	9.88(7.05～13.85)	1.17
4	10.10(9.37～10.82)	9.00(6.09～13.28)	1.10
5	8.90(8.21～9.59)	8.50(5.50～13.15)	0.40
6	7.73(7.08～8.39)	7.47(4.64～12.05)	0.26
7	6.45(5.83～7.07)	7.68(4.59～12.87)	−1.23
8	7.74(7.08～8.39)	7.79(4.49～13.53)	−0.05
9	7.87(7.21～8.54)	8.05(4.49～14.45)	−0.18
10	9.86(9.14～10.58)	9.19(4.96～17.03)	0.67
11	9.41(8.70～10.11)	9.58(5.02～18.29)	−0.17
12	11.01(10.26～11.76)	10.10(5.16～19.89)	0.91

讨　论

人群脑卒中死亡率分布具有季节波动性，大多数传统假设各变量之间是一种线性关系的回归模型，往往不能做出准确的预测。圆分布分析是将具有周期性变化的资料，通过三角函数的变换，使原始数据呈线性资料的一种计算方法，可检验数据在一个周期内的集中趋势，并可计算具体数值[5]。

一定时间内人群脑卒中死亡率是一组随机变量，受多种因素影响，可进行时间序列分析。时间序列模型有多种，较常用的有灰色模型和ARIMA乘积模型。前者对于小样本以及隐含指数函数变化趋势的资料预测具有优势[6]，后者应用局限是样本不能太小($t>20$)，否则预测误差可能会较大；优势是可以综合考虑序列演变的趋势、周期变化和随机干扰因素，借助模型参数的变化对数据进行量化表达，特别适用于找不到预测变量的主要影响因素或者虽然知道影响因素但没有相关数据的预测分析[7]。

ARIMA模型结合圆分布分析可以充分利用两种统计方法的各自优势，判断脑卒中死亡率周期性，获得最佳预测模型。

天津市1999—2005年脑卒中世界标化死亡率的分布总体呈下降趋势，与许多国家研究结果一致。本研究采用实际死亡日期的世界标化月死亡率作为分析时间序列，特别是考虑到人口构成的影响，研究结果增加了模型预测的真实性与可比性。另一方面，天津市1999—2005年脑卒中粗月死亡率相比世界标化月死亡率，下降趋势不明显(表1)。而我国作为发展中国家，未来还将面临快速进入老龄化社会的挑战，天津市脑卒中粗死亡率仍然不容低估。

本研究表明，天津市1999—2005年脑卒中死亡率在冬季(特别是1月份)呈现高水平，有明显的季节性，提示低温可能是重要的影响因素。有研究表明[8]，冬季低温对不同类型脑卒中均增加发病风险。脑卒中死亡的季节分布提示在高发季节应加强监测和医疗资源的配置，同时出台有针对性的预防控制、急诊救治措施，减少脑卒中的死亡。

本研究建立模型的预测值与实际值不完全相同，但实际值大都落入了预测值的95%*CI*，显示模型良好的效力。同时该模型还具有对干预措施效果预测和评估的可扩展性[2]，这对科学考核脑卒中死亡防治措施也将具有重要意义。应注意，脑卒中死亡率时间序列的非平稳性以及观测时间的有限性决定了该模型应为动态模型，今后必须继续收集、不断补充脑卒中死亡信息，才能进一步提高该模型的稳健度、增加预测精度、延长预测周期。

参 考 文 献

[1] 金丕焕.医用统计学. 2版. 上海：复旦大学出版社，2003：211-217.

[2] 徐国强. 统计预测和决策. 2版. 上海：上海财经大学出版社，2005：146-185.

[3] World Health Organization. International classification of diseases, 9th rev. Geneva: WHO, 1977.

[4] World Health Organization. International statistical classification of diseases and related health problems, 10th rev. Geneva: WHO, 1992.

[5] 金永富. 圆分布法分析病毒性肝炎发病的季节性分布. 现代预防医学，2003，30(2)：244-245.

[6] 吴伟，关鹏，郭军巧，等.GM(1,1)灰色模型和ARIMA模型在HFRS发病率预测中的比较研究.中国医科大学学报，2008，37(1)：52-55.

[7] 陈勇，陈建国，朱健，等. 江苏省启东市1972—2001年肺癌发病趋势分析及预测模型比较研究. 中华流行病学杂志，2005，26(12)：955-959.

[8] 刘方，张金良，陆晨. 北京市气温与脑卒中发病关系的时间序列研究. 中华流行病学杂志，2004，25(11)：962-966.

（收稿日期：2008-07-11）

（本文编辑：张林东）