维普资讯 http://www.cqvip.com

重庆医科大学学报 1991年第16卷第2期

三种相关分析计算法的比较研究

数理教研室 杨忠琴 姚 莉
第二医院 石 磊

中图分类号 O159

在医学实验数据处理中，常用相关分析方法是研究分析一个时间移变序列是否相对于时间具有线性或拟线性关系，故一般叫线性相关关系，简称相关关系。常见的10种拟线性关系可见文献[1]。但是这种分析方法能否比较两个数字时间序列或医学信号波形((一般离散化)，特别是两序列本身相对于时间不具有线性或拟线性关系时）之间是否具有线性或拟线性关系呢？一般医学文献中少有提到这个问题。本文将指出对这个问题的肯定回答，并指出主要用于分析两数字时间序列或信号波形之间的相关关系的相关分析方法：模糊子集的相关分析法[2~3]及相关函数分析法[5]。

1．模糊子集相关分析法

任何实验数据序列，都可通过Zadeh映射变换[2]映射到[0,1]的Zadeh空间，构成两个典型的模糊子集，从而用Pal[4]提出的模糊子集的相关分析法计算两序列的相关性。因为医学实验数据大都是离散时间序列，特别是用数字计算机进行计算时，对非离散的医学信号波形，都要首先离散化，以便计算器或计算机进行处理，故以后，我们只提出离散条件下的公式。按Pal的意见，计算两离散模糊子集的相关系数公式如下：

$$C_{\mu_1,\mu_2}=1-\frac{4X_{1,2}}{X_1+X_2} \qquad (1\text{-}1)$$

其中：

$$X_{1,2}=\sum(\mu_1-\mu_2)^2$$

$$X_1=\sum(2\mu_1-1)^2$$

$$X_2=\sum(2\mu_2-1)^2$$

对于单一的数字时间序列，可以看成一种特殊情况，其中一个序列是离散时间序列。

2．两数字时间序列的相关函数分析法

设有两个数字时间序列 $x(n)$，$y(n)$，$n=0,1,2,\cdots,N-1$。要比较两个数字时间(或波形)的相似程度，我们可以把其中一个放大（或缩小）a 倍，然后取它们的误差平方和（或残差平方和）

$$\varepsilon^2=\sum[x(n)-ay(n)]^2 \qquad (2\text{-}1)$$

来衡量其相似程度。根据最小二乘原理，可由(2-1)式算得

$$a=\sum x(n)y(n)/\sum y^2(n) \qquad (2\text{-}2)$$

满足（2-2）式条件下的a值，即具有最小残差平方和为：

$$\varepsilon^2_{min}=\sum x^2(n)-[\sum x(n)y(n)]^2/\sum y^2(n) \qquad (2\text{-}3)$$

用最小残差平方和来估计相似程度的好坏要受数据本身量大小的重大影响，因此按照在文献[6]中提出的估计方法，叫相对残差平方和：

$$\bar{\varepsilon}^2_{min}=\varepsilon^2_{min}/\sum x^2(n) \qquad (2\text{-}4)$$

来估计两序列的相似程度，以便使不同绝对值大小的各序列对之间的相似程度的好坏可以用统一的相对量（百分值）来进行估计。将（2-3）式代入（2-4）式可得：

$$\bar{\varepsilon}^2_{min}=1-[\sum x(n)y(n)]^2/\sum x^2(n)\cdot\sum y^2(n) \qquad (2\text{-}5)$$

维普资讯 http://www.cqvip.com

令

$$\rho_{xy}=\sum x(n)y(n)/[\sum x^2(n)\cdot\sum y^2(n)]^{\frac{1}{2}} \quad (2-6)$$

按许瓦尔兹(Schwartz)不等式

$$|\sum x(n)y(n)|\leqslant[\sum x^2(n)\cdot\sum y^2(n)]^{\frac{1}{2}} \quad (2-7)$$

故

$$|\rho_{xy}|\leqslant 1 \quad (2-8)$$

若记（2-6）式的分子为r_{xy}，分母为$\sqrt{r_{xx}\cdot r_{yy}}$，于是（2-6）式可改写为：

$$\rho_{xy}=\frac{r_{xy}}{\sqrt{r_{xx}\cdot r_{yy}}}$$

当$r_{xy}=\sqrt{r_{xx}\cdot r_{yy}}$时，$\rho_{xy}=1$，这时，$\overline{\varepsilon^2_{min}}=0$，说明$x(n)$和$ay(n)$完全相等，当然它们也最相似。

当$r_{xy}=0$，必有$\rho_{xy}=0$，这时，$\overline{\varepsilon^2_{min}}=1$，说明$x(n)$与$ay(n)$完全不相似。因此，$\rho_{xy}$，$r_{xy}$都可以用来描述$x(n)$和$y(n)$的近似程度，称它们都是数字时间序列$x(n)$和$y(n)$的相关系数。只不过$r_{xy}$体现的是两数字时间序列（或两个固定波形）的相似程度，而ρ_{xy}是归一化了的相关系数，其变化范围在±1之间，因此，通常都使用相关系数ρ_{xy}。在一般情况下，还可使两数字时间序列相对移动，以求得最大的ρ_{xy}，故ρ_{xy}这时可看成时移的函数，故又叫两数字时间序列的相干函数，详见文献[6]。

3．医学统计中常用的相关分析法

一般实验数据处理中常用的相关分析是指一个序列相对于单调递增的时间序列（或其他单调递增的序列）是否具有线性或拟线性关系的分析，我们将指出，经过简单的代数变换，这种方法也可用于估计两个数字时间序列的相关性或波形形状相似性。

在常用统计中，一实验数据序列是否相对于时间具有线性或拟线性关系是指该实验数据序列相对于直线：

$$\hat{y}=a+bx \quad (3-1)$$

的关系紧密或偏离程度。如果实验数据序列具有线性或拟线性关系，则可由最小二乘原理找到适当的a、b，使$\hat{y}_n$与y_n有最小的偏离，即可使残差平方和：

$$Q(a,b)=\sum(y_n-\hat{y}_n)^2=\sum(y_n-a-bx_n)^2 \quad (3-2)$$

最小。经过计算可得：

$$\left.\begin{aligned}b&=\frac{\sum(x_n-\bar{x})(y_n-\bar{y})}{\sum(x_n-\bar{x})^2}\\a&=\bar{y}-b\bar{x}\end{aligned}\right\} \quad (3-3)$$

实验数据序列具有线性或拟线性关系的好坏，可由相对均方差或相对残差均方值及相关系数来衡量：

$$\bar{s}^2=[\sum(y_n-\hat{y}_n)^2/(N-1)]/\sum y_n^2 \quad (3-4)$$

$$r=\frac{\sum(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{\sum(x_n-\bar{x})^2\sum(y_n-\bar{y})^2}}=\frac{L_{xy}}{\sqrt{L_{xx}\cdot L_{yy}}} \quad (3-5)$$

这里值得指出的是：(3-5)式中的变量x并未限定为时间变量，它对于变量x与变量y都是时间的函数，即两个数字时间序列也可以用之计算线性相关系数。因为对于连续函数$x=\varphi(t)$，它的反函数$t=\varphi^{-1}(x)$一定存在，将$t=\varphi^{-1}(x)$代入$y=\psi(t)$中消去t，可得：

$$y=\psi[\varphi^{-1}(x)]=f(x)$$

即y是x的函数，从而可以利用（3-5）式计算其相关系数，也可以拟合它们之间的关系曲线。

4．三种相关系数计算的比较

由（3-5）式与（2-6）式的比较可以看出，在(3-5)式中，令$\bar{x}=0$，$\bar{y}=0$，则(3-5)式与（2-6）式完全一致，因此，可以形式地认为（2-6）式是（3-5）式的特殊情况。

将（1-1）式变成如下形式：

$$C_{\mu_1,\mu_2}=\frac{X_1+X_2-4X_{12}}{X_1+X_2} \quad (4-1)$$

经过计算(4-1)式的分子可以写成：

维普资讯 http://www.cqvip.com

$$8\sum\left(\mu_1-\frac{1}{2}\right)\left(\mu_2-\frac{1}{2}\right)$$

故(4-1)式又可写成

$$C_{\mu_1,\mu_2}=\frac{\sum\left(\mu_1-\frac{1}{2}\right)\left(\mu_2-\frac{1}{2}\right)}{\left[\sum\left(\mu_1-\frac{1}{2}\right)^2+\sum\left(\mu_2-\frac{1}{2}\right)^2\right]2} \quad (4-2)$$

将$\mu_1\in[0,1]$，$\mu_2\in[0,1]$的均值$\bar{\mu}_1$，$\bar{\mu}_2$视为$\frac{1}{2}$，则可以认为（4-2）式的分子与(3-5)式的分子一致。如果将C_{μ_1,μ_2}改写为

$$C_{\mu_1,\mu_2}=\frac{\sum\left(\mu_1-\frac{1}{2}\right)\left(\mu_2-\frac{1}{2}\right)}{\sqrt{\sum\left(\mu_1-\frac{1}{2}\right)^2\cdot\sum\left(\mu_2-\frac{1}{2}\right)^2}} \quad (4-3)$$

则模糊子集相关系数的计算与(3-5)式在形式上是完全一致的。实际数据计算结果表明，(4-3)式与(1-1)式的计算结果基本一致。

比较（3-5）式、(2-6) 式和(4-3)式可以看出，这三个相关系数的表达式，对分别具有0均值的成对数据序列的计算结果完全是一致的。但对非0均值的成对数据序列，要比较两数据序列的线性相关性或两波形相似性，得用（2-6）式的形式，只要两波形完全成比例，则具有相关系数（或相干系数)1。若用（3-5）式或（4-3）式，则只有两序列完全相同时，才可能具有1的相关系数。

另外还需指出的是（3-5）式的一套计算方法还可以拟合出回归方程或叫相关方程，可用以进行对未来事件的预测或推断，(假定未来也符合现在的规律)，而(2-6)式、(1-1) 式或(4-3)式只在于比较线性相关关系，用以进行分类和识别。这是它们在具体运用上的区别。

5．具体计算的比较

这里有两组成对的实验数据序列：小鼠实验性骨折癒合过程中骨痂形成和Ca盐沉积数据；我校第二医院1964～1988年门诊和住院病人的数据，如表1和表2。

表 1　　骨痂形成与Ca盐沉积的实验数据

时　间（天）	5	10	15	20	30
骨痂形成（x）	0.1090	0.1916	0.1060	0.0700	0.0440
$\mu_1(x)$	0.569	1.000	0.553	0.365	0.230
Ca盐沉积（y）	3.5060	4.4960	3.4870	2.8765	2.0230
$\mu_2(y)$	0.780	1.000	0.776	0.640	0.450

表 2　　64～88年门诊人数与住院人数数据

年　份　（年）	64	65	66	67	68	69	70	71
门诊人数（x）	439438	458008	542238	500882	248542	378379	497903	380913
$\mu_1(x)$	0.665	0.693	0.821	0.758	0.376	0.573	0.754	0.576
住院人数(y)	5498	6552	7347	6257	4823	6108	6349	6519
$\mu_2(y)$	0.673	0.802	0.900	0.766	0.591	0.748	0.777	0.798
年　份　（年）	72	73	74	75	76	77	78	79
门诊人数（x）	366279	359564	322910	378282	340471	421973	380513	395267
$\mu_1(x)$	0.554	0.544	0.489	0.573	0.515	0.639	0.576	0.598
住院人数(y)	5952	5230	5411	5277	3772	3846	3866	5142
$\mu_2(y)$	0.729	0.640	0.663	0.646	0.462	0.471	0.473	0.630

维普资讯 http://www.cqvip.com

续表

年　份　(年)	80	81	82	83	84	85	86	87	88
门诊人数(x)	557978	630260	651557	660736	654140	601208	568724	577250	534448
$\mu_1(x)$	0.844	0.954	0.986	1.000	0.990	0.910	0.861	0.874	0.809
住院人数(y)	7724	8167	8107	7998	7331	6447	4869	5506	5741
$\mu_2(y)$	0.946	1.000	0.993	0.979	0.898	0.789	0.596	0.674	0.703

1．对表1数据的计算结果

$C_{\mu_1,\mu_2}=0.736$　　按（1-1）式计算

$C_{\mu_1,\mu_2}=0.722$　　按（4-3）式计算

$\rho_{xy}=0.976$　　按（2-6）式计算

$r_{xy}=0.970$　　按（3-5）式计算

2．对表2数据的计算结果

$C_{\mu_1,\mu_2}=0.89286$　　按（1-1）式计算

$C_{\mu_1,\mu_2}=0.89287$　　按（4-3）式计算

$\rho_{xy}=0.98498$　　按（2-6）式计算

$r_{xy}=0.70604$　　按（3-5）式计算

此外，我们还对表1、表2的数据分别用多项式拟合的办法进行了拟合。

1．骨痂形成与时间的关系方程

$$x(t)=-0.5360871+0.221357t-0.0226737t^2+9.053296\times10^{-4}t^3-1.246685\times10^{-5}t^4$$

$r=0.9977772$　　$s=6.487739\times10^{-3}$

2．Ca盐沉积与时间的关系方程：

$$y(t)=-3.767123+2.47343t-0.2491886t^2+9.771968\times10^{-3}t^3-1.333157\times10^{-4}t^4$$

$r=0.9990388$　　$s=2.122996\times10^{-3}$

3．骨痂形成与Ca盐沉积之间的关系方程：

$$y=1.6296+15.8288x$$

$r=0.9703$

4．门诊人数与时间的关系方程

$$x(t)=610289.5-77077.91t+6784.064t^2-149.974t^3$$

$r=0.781729$　　$s=0.1825729$

5．住院人数与时间的关系方程

$$y(t)=652599-630.9757t+227.962t^2-19.80819t^3-0.886863t^4+0.1196864t^5-4.63943\times10^{-4}t^6-7.838689\times10^{-5}t^7-5.553527\times10^{-6}t^8+2.0435\times10^{-7}t^9$$

$r=0.8358386$　　$s=0.1337838$

6．门诊人数与住院人数之间的关系方程：

$$y=2342.437+0.0077x$$

$r=0.706$

参考文献

[1] 谢正祥，等译。生物学中BASIC微计算机模型，科技文献出版社重庆分社，重庆1987；61

[2] 杨忠琴，生物医学工程学杂志，1989；7(1)：55

[3] 粟载福、杨忠琴等。模糊数学与医学，科技文献出版社重庆分社，重庆1989；368

[4] Pal, S.K.et al. Fuzzy mathematical Approach to Pattern Recognition, WE NEW Delhi, 1986；64

[5] 宗孔德等。数字信号处理，清华大学出版社，北京1988；118

[6] 何振亚。数字信号处理的理论与应用(下)，人民邮电出版社，北京1987；194

(收到日期1990-09-17)

维普资讯 http://www.cqvip.com

重庆医科大学学报 1991年第16卷第2期

分子生物学技术在临床病理学中的应用

病毒性肝炎研究所 范 淳
病理生理教研室 范维珂

中图分类号 Q7

近年来，分子生物学和免疫学的发展为临床病理学提供了新的手段。重组DNA技术应用于临床病理学，代表了最新的基础研究向临床实验室的过渡，是特殊染色，酶组织化学，电镜，单克隆抗体，免疫组织化学及流式细胞计数器等技术的延伸和发展，构成了目前病理学诊断与研究中最为先进而有力的工具。特别是分子探针技术在研究组织中基因的结构及表达，绘制基因图，微生物病原体及肿瘤基因的诊断以及DNA分子重排的分子特征方面起着极为重要的作用[1]。而更为晚近所产生的聚合酶链反应（简称PCR）[2]，又大大提高了检测的灵敏度，从而将分子生物学研究推向了一个更新的高度。毫无疑问，研究基因的结构和表达与组织形态学，临床表现及预后的关系，将进一步提高我们对人类疾病的诊断与治疗的能力。因此，本文拟对可用于临床实验室的分子生物学技术及其应用作一简要的介绍。

一、DNA的结构与来源

重组DNA技术最大的优点就是其特异性与敏感性。目前，临床实验室最常用的技术有二种：（1）分子杂交试验；（2）限制性内切酶分析及疾病相关基因多态性分析。

DNA是由二条核苷酸链相互缠绕而形成的双螺旋结构。每条链的核苷酸碱基都特异性地与其相对链的互补碱基通过氢键相连。胸腺嘧啶与腺嘌呤相连，胞嘧啶与鸟嘌呤相连。加热时，碱基间的氢键发生断裂，二个碱基分离，称为DNA变性。而当冷却时，二

COMPARING INVESTIGATIONS FOR CALCULATING METHOD ON THE THREE CORRELATION ANALYSIS

Yang Zhongqin
Dept. of Math. -phys.

In this paper, the relationship between calculating methods of the three correlation analyses and the differences in their applications were presented. Another expression about correlation analysis between fuzzy sets was suggested. Examples in calculating the practical datum on digital-time sequences with three correlation analysis were given

KEY WORDS CORRELATION ANALYSIS; DIGITAL-TIME SEQUENCE; FUZZY SETS; WAVEFORM SIMILARITY