相关系数显著性检验的几何意义

姚菊香[1],王盘兴[1],鲍学俊[2],卢楚翰[1]
(1.南京信息工程大学 大气科学学院,江苏 南京 210044; 2.上海市气象信息传媒中心,上海 200030)

摘要:从几何学角度阐明了相关系数显著性检验的意义。对于来自正态分布的样本,利用其距平序列对应的随机向量在高维空间中均匀分布的性质,在母体无相关假定下,用几何方法求得了显著性水平 α和样本容量 n下的临界相关系数 $r'_{\alpha,n}$的表达式,并验证了它等于由 t分布求得的临界相关系数 $r_{\alpha,n}$,从而给出了相关系数显著性检验的直观理解。

关键词:相关系数;显著性检验;几何意义
中图分类号:O212.1 **文献标识码**:A **文章编号**:1000-2022(2007)04-0566-05

Geometric Meaning of the Significance Test of Correlation Coefficient

YAO Ju-xiang[1], WANG Pan-xing[1], BAO Xue-jun[2], LU Chu-han[1]
(1. School of Atmospheric Sciences, NUIST, Nanjing 210044, China;
2. Shanghai Meterological Media Center, Shanghai 200030, China)

Abstract: Analysis of correlation coefficient is widely used in the study of short-term climate variation and prediction. The meaning of the significance test of correlation coefficient is elucidated from geometric angle. Based on the character that the stochastic vectors corresponding to the anomaly sequence of the samples with a normal distribution, uniformly distribute in the high dimension space, supposing that the samples come from independent parent populations, the expression of critical coefficient $r'_{\alpha,n}$ under the conditions of significance level α and sample capacity n was obtained by the methods of geometry. That the $r'_{\alpha,n}$ equals to the critical correlation coefficient $r_{\alpha,n}$ obtained from t-distrbution was validated. So the intuitive understanding of the significance test of correlation coefficient is given.

Key words: correlation coefficient; significance test; geometric meaning

0 引言

在短期气候变化及其预测研究领域,由于研究对象间关系的复杂性,大量分析工作仍只能在线性相关的层面进行[1-4]。气象上,在研究青藏高原地表温度对华北汛期降水变化的影响[5]、中国西南汛期降水的振动和分布及其与印度洋海温异常的关系[6]、华北夏季降水与哈得孙湾海冰的关系[7]、西北地区东部夏季温度特征及与热带 SSTA 的相关关系[8]时,都用到线性相关的分析方法。在线性相关分析中,相关系数的显著性检验是据样本相关系数 r估计母体相关与否的方法。对此,统计学教科书[9-11]已有严格论证和系统介绍。王盘兴等[12-13]将相关系数 r表示成 n维欧氏空间(记为 E^n)中两个向量间夹角的余弦,从几何学的角度改写了 r的表达式,阐明了 r的几何意义。研究表明,从几何学角度观察、处理一系列气象学中的问题[14-17],如两个场之间相似性问题[16],较之从统计学角度更为明晰、简洁。本文从几何学角度分析并阐明相关系数显著性检验的意义。这种分析有助于对气象统计学问题的理解,从而提高解决短期气候异常分析和预测实际问题的能力。

1 r的几何实质

由文献[12-13]可知,随机变量 X, Y的一对容量为 n的样本

收稿日期:2006-06-20;**改回日期**:2007-03-29
基金项目:国家自然科学基金重点资助项目(40633018)
作者简介:姚菊香(1980-),女,江苏通州人,助教,硕士,研究方向:短期气候预测,jxyao@126.com.

$$x_i, y_i, i = 1, 2, \cdots, n, \tag{1}$$

可看作 E^n 中的一对随机向量

$$x = (x_1\ x_2\ \cdots x_n), y = (y_1\ y_2\ \cdots y_n)。 \tag{2}$$

根据文献[18],它们可分解为

$$x = \bar{x} + x', \quad y = \bar{y} + y'。 \tag{3}$$

式中:均值向量 $\bar{x}, \bar{y}$ 的每个分量为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i。 \tag{4}$$

距平向量 x', y' 的每个分量为

$$x_i' = x_i - \bar{x}, \quad y_i' = y_i - \bar{y}。 \tag{5}$$

由文献[12]可知,$(\bar{x}, x') = 0$, $(\bar{y}, y') = 0$,故 $x' \perp \bar{x}$, $y' \perp \bar{y}$。又由于 $\bar{x}, \bar{y}$ 与 E^n 中的幺向量 $e = (1\ 1\ \cdots 1)$ 共线,故 x', y' 落在 E^n 中垂直于 e 的子空间 E^{n-1}($n-1$维空间)中;它们的交角 $\theta = \langle x', y' \rangle$ 的余弦为 x, y 的相关系数,即

$$r = \cos\langle x', y' \rangle。 \tag{6}$$

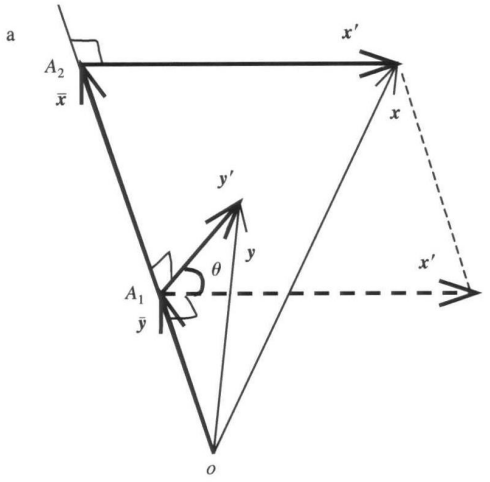
定义标准化向量

$$\tilde{x}' = \frac{x'}{\|x'\|}, \quad \tilde{y}' = \frac{y'}{\|y'\|}。 \tag{7}$$

标准化使 $\|\tilde{x}'\| = \|\tilde{y}'\| = 1$,故 $\tilde{x}', \tilde{y}'$ 的矢端均落在 E^{n-1} 中单位半径球面(其维数为 $n-2$)上;而标准化不改变 x', y' 的方向,故得 r 的另一表达式:

$$r = \cos\langle \tilde{x}', \tilde{y}' \rangle。 \tag{8}$$

图 1分别给出了 E^n 中 x, y 的分解及 E^{n-1}(E^{n-2})中 x', y'($\tilde{x}', \tilde{y}'$)的交角 θ,θ 的余弦即为 r。



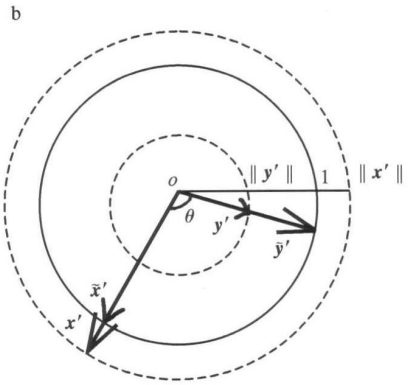


图 1　相空间中 x, y 的分解的几何示意图(折线:向量垂直;弧线:向量交角)

a. E^n; b. E^{n-1}(E^{n-2})

Fig. 1　Geometric graph for the decomposition of x and y in phase space (polyonal line: right-angle; arc: angle of intersection of vectors)

a. E^n; b. E^{n-1}(E^{n-2})

2　r显著性检验的统计意义

按文献[9],当样本(1)来自于正态无相关母体 X, Y时,其样本相关系数 r的概率密度函数为

$$f(r) = \frac{n-2}{\pi}(1-r^2)^{\frac{n-4}{2}}\int_0^1 z^{-\frac{1}{2}}(1-z)^{\frac{n-3}{2}}dz。 \tag{9}$$

作变换

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}, \tag{10}$$

可得 t的概率密度函数为

$$g(t) = \frac{1}{\sqrt{n-2}}\cdot\frac{1}{B\left(\frac{n-2}{2}, \frac{1}{2}\right)}\cdot\frac{1}{\left(1+\frac{t^2}{n-2}\right)^{\frac{n-1}{2}}}, \tag{11}$$

则 t服从自由度为 $n-2$ 的学生氏 t分布[9]。在给定显著性水平 α(它通常是一个小量,如 $\alpha = 0.01, 0.05$)下,针对样本容量 n,可由 g的概率分布函数 $P(|t| \geq t_{\alpha,n}) = \alpha$ 计算出 t的临界值 $t_{\alpha,n}$,并据(10)式求得 r的临界值

$$r_{\alpha,n} = \frac{t_{\alpha,n}}{\sqrt{(n-2) + t_{\alpha,n}^2}}, \tag{12}$$

它使

$$P(|r| \geq r_{\alpha,n}) = \alpha。 \tag{13}$$

因为 α 是小量,故在一次试验中出现 $|r| \geq r_{\alpha,n}$ 是小概率事件。根据实际推断原理[10],若在一次试验中出现 $|r| \geq r_{\alpha,n}$,则可在显著性水平 α 下否定"样

本 (1)来自无相关母体 "的假设,判断母体相关;反之,若 $|r| < r_{\alpha,n}$,则肯定"样本 (1)来自无相关母体 ",判断母体不相关。

3 r显著性检验的几何意义

从几何角度看,当样本 (1)来自正态母体 X, Y时,随机向量 $x'(y')$是 E^{n-1}中各向同性的向量。$x'(y')$各向同性是指它出现在 E^{n-1}中所有方向上的概率相等,$\tilde{x}'(\tilde{y}')$矢端均匀出现在 E^{n-1}中单位半径超球球面 (它的维数为 n - 2)上。图 2给出了 n = 3时,由 N(0, 1)随机试验给出的 100个 x和 $\tilde{x}$的图像,它们直观显示了上述几何性质。

当正态母体 X, Y无相关时,$x'(\tilde{x}')$与 $y'(\tilde{y}')$取向 (位置)无关。若在 E^{n-1}中取 x为极轴,$\tilde{x}$即为 E^{n-1}中单位半径超球球面的北极点;y在 E^{n-1}中均匀取向,$\tilde{y}$在 E^{n-1}中单位半径超球球面上均匀取位。对于给定的小量 α及样本容量 n,上述几何均匀性将使 $\tilde{y}$出现在单位半径超球球上 $\theta \le \theta'_\alpha$的北极区或 y出现在 E^{n-1}中 $\theta \in [0, \theta'_\alpha]$角域内的概率,等于 E^{n-1}中 $\theta \le \theta'_\alpha$的极冠区单位半径球面面积 $(S_{\theta'_\alpha})_{n-1}$与半球球面面积 $\left(S_{\frac{\pi}{2}}\right)_{n-1}$之比,即

$$P(0 \le \theta \le \theta'_\alpha) = \frac{(S_{\theta'_\alpha})_{n-1}}{(S_{\pi/2})_{n-1}} = \alpha。 \quad (14)$$

上式已考虑了 E^{n-1}中超球对 $\theta = \pi/2$的对称点,只给出了超球"北半球 "部分。因为 α是一个小量,故"y落在 $\theta \in [0, \theta'_{\alpha,n}]$ "或"$\tilde{y}$落在 $\theta \le \theta'_{\alpha,n}$的极冠区 "是一个小概率事件;按实际推断原理,可据它在一次抽样中出现与否,判断正态母体 X, Y是否相关。这是 r显著性检验的几何意义。

4 验证

对于给定的 α和 n, (12)式给出了相关系数的统计学临界值 $r_{\alpha,n}$;由 (14)式确定的 $\theta'_{\alpha,n}$可以给出其几何学临界值 $r'_{\alpha,n}$。问题的关键归结为

$$r'_{\alpha,n} = r_{\alpha,n} \quad (15)$$

是否成立。式中:

$$r'_{\alpha,n} = \cos\theta'_{\alpha,n}。 \quad (16)$$

为了验证 (15)式,首先需要对给定的 α、n求出 $\theta'_{\alpha,n}$和 $r'_{\alpha,n}$。由文献 [19], E^{n-1}中与 $\theta'_{\alpha,n}$对应的单位半径超球极冠区面积为

$$(S_{\theta'_\alpha})_{n-1} = \begin{cases} \dfrac{2^{\frac{n-1}{2}} \cdot \pi^{\frac{n-3}{2}}}{(n-4)!!} \displaystyle\int_0^{\theta'_\alpha} \sin^{n-3}\theta d\theta, & n\text{为奇数}; \\ \dfrac{\pi^{\frac{n-2}{2}} \cdot 2^{\frac{n-2}{2}}}{(n-4)!!} \displaystyle\int_0^{\theta'_\alpha} \sin^{n-3}\theta d\theta, & n\text{为偶数}。 \end{cases} \quad (17)$$

半球面积为

$$\left(S_{\frac{\pi}{2}}\right)_{n-1} = \begin{cases} \dfrac{1}{2} \cdot \dfrac{(2\pi)^{\frac{n-1}{2}}}{(n-3)!!}, & n\text{为奇数}; \\ \dfrac{1}{2} \cdot \dfrac{\pi^{\frac{n-2}{2}} \cdot 2^{\frac{n}{2}}}{(n-3)!!}, & n\text{为偶数}。 \end{cases} \quad (18)$$

将 (17)、(18)式代入 (14)式,得积分方程

$$\int_0^{\theta'_\alpha} \sin^{n-3}\theta d\theta = \begin{cases} \dfrac{\alpha\pi(n-4)!!}{2(n-3)!!}, & n\text{为奇数}; \\ \dfrac{\alpha(n-4)!!}{(n-3)!!}, & n\text{为偶数}。 \end{cases} \quad (19)$$

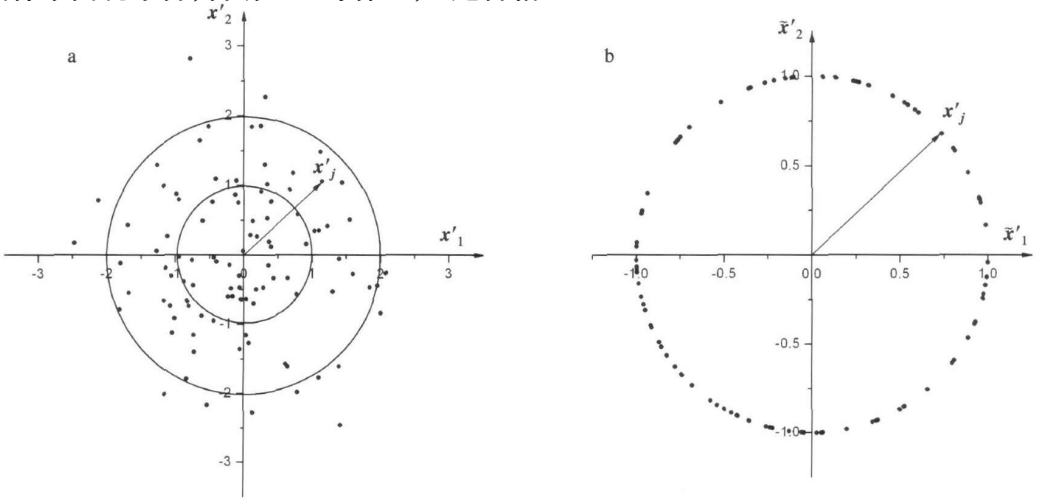


图 2 由 N(0, 1)随机试验产生的 100个 x'(a)、$\tilde{x}'$(b) (n = 3)

Fig. 2 100 (a) x's and (b) $\tilde{x}'$s produced by the N(0, 1) random test

式中:!! 为双阶乘号,并规定 0!! =1!! =1, 2!! =2。对 α=0.05、0.01和一些 n值,从积分方程(19)中解得 $\theta'_{\alpha,n}$,并据(16)式求得 $r'_{\alpha,n}$,结果列于表 1。

表 1　几何分析求得的 $\theta'_{\alpha,n}$及 $r'_{\alpha,n}$

Table 1　$\theta'_{\alpha,n}$ and $r'_{\alpha,n}$ obtained from Equations (19) and (16), respectively

n	α=0.05		α=0.01	
	$\theta'_{\alpha,n}$/(°)	$r'_{\alpha,n}$	$\theta'_{\alpha,n}$/(°)	$r'_{\alpha,n}$
3	4.50	0.996 9	0.90	0.999 9
4	18.19	0.950 0	8.11	0.990 0
5	28.56	0.878 3	16.52	0.958 7
7	41.02	0.754 5	29.01	0.874 5
10	50.81	0.631 9	40.13	0.764 6
20	63.65	0.443 8	55.85	0.561 4
50	73.82	0.278 7	68.84	0.361 0
100	78.66	0.196 6	75.14	0.256 5
200	82.02	0.138 8	79.53	0.181 8
500	84.97	0.087 7	83.39	0.115 1
1 000	86.45	0.062 0	85.33	0.081 4

经与统计途径求得的相关系数临界值 $r_{\alpha,n}$[9]比较,(15)式成立。r显著性检验的几何意义得到验证。

对 α=0.05和 n=3、4,图 3直观地给出了 r显著性检验的几何意义。对 n=3(图 3a),$\tilde{y}$应在半圆弧$\overleftarrow{CAC}$上均匀取位;若 x, y无相关,$\frac{\overleftarrow{BAB}'_{CAC}}{}$=0.05,由此求得 $\theta'_{0.05,3}$=4.5°, $r'_{0.05,3}$=0.996 9=$r_{0.05,3}$。对 n=4(图 3b),$\tilde{y}$应在 A为北极点的半球球面上均匀取位;若 x, y无相关,由弧$\overleftarrow{AB}$绕 x一周构成的极冠区面积与该半球面积之比$\frac{(S_{0.05})_3}{(S_{\frac{\pi}{2}})_3}$=0.05,求得 $\theta'_{0.05,4}\approx 18.2°$, $r'_{0.05,4}$=0.950 0=$r_{0.05,4}$。

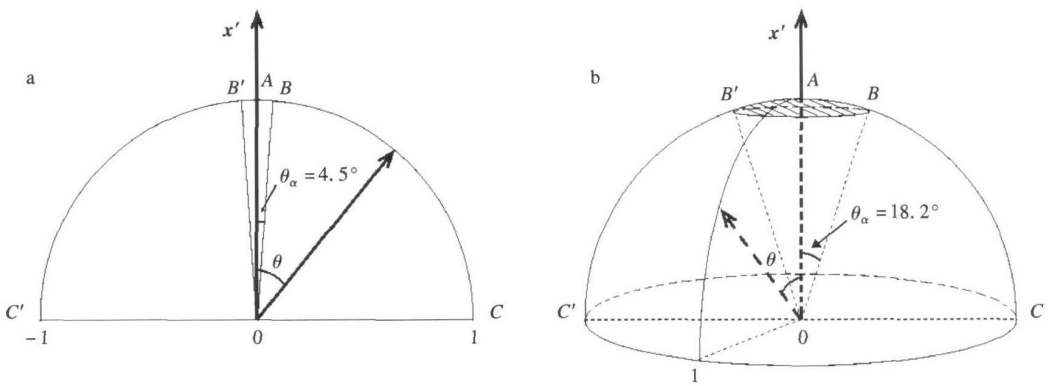


图 3　r显著性检验几何意义　　a. n=3; b. n=4

Fig. 3　Geometric meaning of the significance test of r　　a. n=3; b. n=4

5　小结

相关系数分析在短期气候变化及预测研究中有着广泛应用,本文从几何学角度阐明了相关系数显著性检验的意义。对来自正态分布的样本,利用其距平序列对应的随机向量在高维空间中散布均匀的性质,在母体无相关假定下,用几何方法求得了显著性水平 α和样本容量 n下的临界相关系数 $r'_{\alpha,n}$的表达式,并验证了它等于由 t分布求得的临界相关系数 $r_{\alpha,n}$,从而给出了相关系数显著性检验的直观理解。

致谢:吴诚鸥教授对此文提出了宝贵意见,谨致谢忱!

参考文献:

[1] 赵宗慈,高学志,罗勇,等. 气候模式作年季预报的几个问题[R]. LASG, Technical Report, 1997, 3: 78-90.

[2] 王艳玲,郭品文. 春季北方气旋活动的气候特征及与气温和降水的关系[J]. 南京气象学院学报, 2005, 28(3): 391-397.

[3] 陈兵,郭品文,向淦川. 夏季低空越赤道气流与 ENSO的关系[J]. 南京气象学院学报, 2005, 28(1): 36-43.

[4] Li Chongyin, Zhou Wen, Jia Xiaolong, et al. Decadal/Interdecadal variation of the ocean temperature and its impacts on climate[J]. Adv Atmos Sci, 2006, 23(6): 964-981.

[5] 余锦华,荣淑艳,任健. 青藏高原地表温度对华北汛期降水变化的影响[J]. 气象科学, 2005, 25(6): 579-586.

[6] 晏红明,肖子牛. 中国西南汛期降水的振动和分布及其与印度

洋海温异常的关系 [J]. 气象科学, 2001, 21(1): 54-63.
[7] 谢付莹, 何金海. 华北夏季降水与哈得孙湾海冰的相关分析 [J]. 南京气象学院学报, 2003, 26(3): 308-316.
[8] 王兰宁, 田武文, 黄祖英, 等. 西北地区东部夏季温度特征及与热带 SSTA 的相关关系 [J]. 气象科学, 2000, 20(1): 23-29.
[9] Fisz M. 概率论及数理统计 [M]. 王福保, 译. 上海: 上海科学技术出版社, 1962.
[10] 复旦大学数学系. 概率论与数理统计 [M]. 上海: 上海科学技术出版社, 1961.
[11] 盛骤, 谢式千, 潘承毅. 概率论与数理统计 [M]. 北京: 高等教育出版社, 1989.
[12] 王盘兴, 李丽平, 周春华. 气象统计若干基本问题的几何实质 (I) [J]. 山东气象, 2002, 22(4) 4: 3-7.
[13] 王盘兴, 李丽平, 周春华. 气象统计若干基本问题的几何实质 (II) [J]. 山东气象, 2003, 23(1): 8-12.
[14] 王盘兴. 论变量线性相关系数的相互制约及其对多元线性回归方程拟合优度的影响 [J]. 气象学报, 1986, 44(1): 70-81.
[15] Wang Panxing. A map of local pattern analogue coefficient: A tool for displaying circulation anomaly [J]. Acta Meteor Sinica, 1992, 6(2): 325-331.
[16] 王盘兴, 李刚, 王建新, 等. 原观测场时间序列两个统计量的相似性讨论 [J]. 气象学报, 1998, 56(6): 764-751.
[17] 王盘兴, 李丽平, 周伟灿. 某些气象统计学问题的几何学分析 [J]. 气象教育与科技, 2001, 23(1): 1-5.
[18] 洛伦茨. 大气环流的性质和理论 [M]. 北京大学地球物理系气象专业, 译. 北京: 科学出版社, 1976.
[19] 华东师范大学数学系. 数学分析: 下册 [M]. 3版. 北京: 高等教育出版社, 2002.