相关系数检验表的新算法^{*}

周永宏 郑大伟

(中国科学院上海天文台, 上海 200030)

提 要

应用 Monte Carlo 方法, 通过大量的数值计算, 得到一个容量大、稳定性好的相关系数临界值表, 供大家参考使用。

主题词: 相关系数 — Monte Carlo 方法

一、 引 言

相关分析技术是研究变量间关系的重要统计方法, 已经成为多种研究领域常用的数据处理工具; 而对其统计结果显著水平的检验也同样极为重要^[1,2]。然而, 大多数理论统计书籍附有的相关系数检验表包括的范围较小, 使得难以检验较大样本间相关结果的可靠性。所以, 我们决定借助 Monte Carlo 方法, 通过大量的数值计算, 得到一个较大容量的相关系数临界值表, 以满足我们天文工作上的实际需要, 同时也可供大家参考使用。

二、 应用 Monte Carlo 方法估计相关系数的临界值

在信号或数据的分析中, 相关系数是度量两列静态测量数据相关性的定量指标。设有两个时间序列 $x_1(t)$ 和 $x_2(t)$, 其相关系数 ρ 可由下式定义^[3]:

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} \quad (1)$$

其中 σ_{11} 和 σ_{22} 分别是 $x_1(t)$ 和 $x_2(t)$ 的方差, σ_{12} 是它们之间的协方差。

相关检验理论涉及的内容较多(参看文献[1~7]), 这里只简要指出, ρ 是个随机变量, 若 x_1 和 x_2 都服从正态分布, 根据零假设 H_0 (两者无相关性) 时 ρ 的分布函数, 可以制作一个 ρ 分布的数值表。给定置信度 α , 可以根据自由度 dof, 确定相应的临界值 $\rho_{1-\alpha}$, 使概率

$$P(|\rho| > \rho_{1-\alpha}) = \alpha \quad (2)$$

当由样本算得的 $|\rho| > \rho_{1-\alpha}$ 时, 以置信度 α 否定零假设 H_0 , 认为相关性是显著的; 反之, 则相关性不显著。

Monte Carlo 方法是一类通过随机变量的统计试验和随机模拟,求解数学、物理和工程技术问题近似解的数值方法^[8~10]。它的基本思想是为了求解数学、物理和工程技术等方面的问题,首先建立一个概率模型或随机过程,使它的参数等于问题的解,然后通过对模型或过程的观察或抽样试验来计算所求参数的统计特征,最后给出所求解的近似值。而解的精度可以用估计值的标准离差来表示。

近年来电子计算机运算速度的飞速提高,促使 Monte Carlo 方法得到更加深入广泛的应用。本文将采用 Monte Carlo 方法来估计相关系数的临界值,具体步骤如下:(a) 产生一对 n 点独立、正态分布的随机序列^[11],相应的自由度 dof 为 $n - 2$,然后按照公式(1)计算它们间的相关系数 ρ 。(b) 重复步骤(a) 5000 次,可以得到 5000 个 ρ 。接着将它们按从小到大的顺序排列,并找出第 4500,4750 和 4950(分别对应于总数 5000 的 90%,95%和 99%)个 ρ ,记为 ρ_{90} , ρ_{95} 和 ρ_{99} 。(c) 重复上述步骤 10 次,依此计算出 ρ_{90} , ρ_{95} 和 ρ_{99} 的平均值和标准离差,其平均值即可作为相应于给定的 dof、置信度 $\alpha = 0.10, 0.05$ 和 0.01 的相关系数临界值。改变 dof 的数值,计算相应的相关系数临界值及其标准离差,结果分别列于表 1 和表 2。因为对于每个 dof,我们都赋予不同的初值来生成随机序列,所以表 1 中的相关系数临界值应彼此独立。

表 1 相关系数临界值表

Table 1 A table of threshold values of correlation coefficients

dof α	0.10	0.05	0.01	dof α	0.10	0.05	0.01
1	0.988	0.997	1.000	80	0.183	0.218	0.283
2	0.901	0.951	0.990	90	0.171	0.204	0.266
3	0.806	0.879	0.958	100	0.162	0.193	0.254
4	0.728	0.809	0.916	120	0.150	0.178	0.233
5	0.669	0.755	0.874	140	0.138	0.165	0.216
6	0.623	0.709	0.835	160	0.130	0.154	0.201
7	0.582	0.668	0.797	180	0.123	0.147	0.193
8	0.547	0.633	0.766	200	0.116	0.138	0.180
9	0.519	0.598	0.733	220	0.111	0.132	0.171
10	0.496	0.572	0.702	240	0.106	0.126	0.165
11	0.474	0.552	0.685	260	0.102	0.122	0.160
12	0.458	0.534	0.663	280	0.098	0.116	0.153
13	0.443	0.516	0.640	300	0.095	0.113	0.148
14	0.426	0.501	0.624	320	0.092	0.110	0.146
15	0.414	0.480	0.605	340	0.089	0.107	0.139
16	0.399	0.469	0.587	360	0.087	0.104	0.135
17	0.389	0.456	0.573	380	0.084	0.101	0.132
18	0.377	0.443	0.561	400	0.083	0.098	0.128
19	0.369	0.433	0.550	450	0.077	0.093	0.121
20	0.360	0.424	0.542	500	0.073	0.087	0.114
21	0.352	0.413	0.526	550	0.070	0.084	0.110
22	0.342	0.401	0.508	600	0.067	0.080	0.105
23	0.337	0.396	0.506	650	0.065	0.077	0.100
24	0.328	0.388	0.496	700	0.062	0.074	0.095
25	0.324	0.383	0.487	750	0.060	0.071	0.094
26	0.317	0.373	0.477	800	0.058	0.069	0.091
27	0.311	0.366	0.469	850	0.056	0.067	0.088
28	0.305	0.359	0.459	900	0.055	0.066	0.087
29	0.302	0.355	0.456	950	0.053	0.063	0.083
30	0.295	0.347	0.446	1000	0.052	0.062	0.082
35	0.274	0.324	0.418	1500	0.043	0.051	0.066
40	0.257	0.305	0.390	2000	0.037	0.044	0.058
50	0.231	0.274	0.351	3000	0.030	0.036	0.047
60	0.210	0.249	0.323	4000	0.026	0.031	0.041
70	0.195	0.233	0.303	5000	0.023	0.028	0.036

表 2 相关系数临界值的标准离差表

Table 2 A table of standard deviations of correlation coefficient threshold values

dof	α	0.10	0.05	0.01	dof	α	0.10	0.05	0.01
1		0.001	0.000	0.000	80		0.002	0.002	0.003
2		0.005	0.004	0.002	90		0.002	0.003	0.005
3		0.005	0.004	0.004	100		0.002	0.002	0.006
4		0.007	0.008	0.006	120		0.001	0.002	0.006
5		0.006	0.005	0.007	140		0.002	0.003	0.004
6		0.007	0.005	0.008	160		0.002	0.002	0.005
7		0.007	0.008	0.003	180		0.001	0.002	0.005
8		0.008	0.007	0.007	200		0.001	0.002	0.003
9		0.004	0.005	0.008	220		0.001	0.001	0.003
10		0.005	0.007	0.006	240		0.001	0.001	0.003
11		0.004	0.005	0.007	260		0.001	0.002	0.002
12		0.006	0.008	0.006	280		0.001	0.001	0.002
13		0.007	0.010	0.010	300		0.001	0.002	0.003
14		0.005	0.007	0.006	320		0.001	0.002	0.003
15		0.005	0.005	0.009	340		0.001	0.002	0.003
16		0.003	0.005	0.008	360		0.001	0.002	0.003
17		0.004	0.004	0.009	380		0.001	0.001	0.002
18		0.004	0.004	0.008	400		0.002	0.002	0.002
19		0.005	0.004	0.009	450		0.001	0.001	0.003
20		0.004	0.005	0.008	500		0.000	0.001	0.001
21		0.004	0.004	0.009	550		0.001	0.001	0.002
22		0.005	0.005	0.007	600		0.001	0.001	0.002
23		0.006	0.004	0.007	650		0.001	0.001	0.001
24		0.004	0.004	0.009	700		0.001	0.001	0.001
25		0.003	0.004	0.008	750		0.000	0.001	0.002
26		0.006	0.005	0.008	800		0.001	0.001	0.002
27		0.004	0.004	0.007	850		0.000	0.001	0.002
28		0.003	0.005	0.005	900		0.001	0.000	0.001
29		0.004	0.006	0.010	950		0.001	0.001	0.002
30		0.003	0.005	0.008	1000		0.000	0.001	0.002
35		0.004	0.004	0.008	1500		0.000	0.000	0.001
40		0.004	0.004	0.006	2000		0.000	0.000	0.001
50		0.003	0.005	0.006	3000		0.000	0.000	0.001
60		0.002	0.003	0.008	4000		0.000	0.000	0.001
70		0.002	0.003	0.005	5000		0.000	0.000	0.001

从表2可以看出,本文通过MonteCarlo方法获得的相关系数临界值波动范围仅在 ± 0.01 以内,并且其相应值与文献[1]的附表17符合很好,两者相差不超过0.01。这说明我们的运算量已经足够大,因而结果稳定性也相当好。

根据表1数据可以绘成图1,自上而下的三根曲线分别是对应于置信度 $\alpha = 0.01, 0.05$ 和0.10的相关系数临界曲线。对于某些所需的dof、而表1中未列出,可以很方便地通过线性插值法得到相应的相关系数临界值。

三、应用实例

由于宇宙空间、地球表层及其内部等多种物理因素的影响,地球自转速率包含着时间尺度为几天到地质年代时间跨度的变化过程^[12-13]。地球自转速率在几年时间尺度上的波动通常用日长的年际变化来表征。我们采用1964年7月到1993年12月日长变化的月平均资料,让它

图 1 相关系数的临界值曲线,自上而下分别对应于置信度水平 $\alpha = 0.01, 0.05$ 和 0.10 。

Fig. 1 The threshold value curves of correlation coefficients, which from the upper to the bottom correspond respectively to the confidence levels $\alpha = 0.01, 0.05$, and 0.10 。

通过多级数字滤波器^[14],滤掉 2 年以下的高频分量和 10 年以上的长期起伏,这样便得到几年时间尺度、非准周期性的日长年际变化序列,绘于图 2 的上部。

图 2 日长年际变化序列和南方涛动指数

Fig. 2 Interannual length-of-day variation series and southern oscillation index

南方涛动突出表现为热带太平洋地区东西半球大气的一种不规则涛动,El Nino 现象伴随于南方涛动过程中,且略迟后 2~3 个月. 在 El Nino 现象发生时,一股温度比平均值高出 2℃ 的水系横穿热带太平洋地区,使秘鲁沿岸附近的海水温度显著上升,从而引起热带太平洋地区和全球大部分地区的气象异常. 由于 El Nino 现象是南方涛动的一个阶段,所以常常把它们合称为 El Nino/南方涛动(ENSO)现象^[15]. ENSO 的强度通常用南方涛动指数(SOI)来描

述,它是基于塔希堤(Tahiti,18°S,148°W)和达尔文(Darwin,12°S,131°E)两地季节性调整后的海平面气压差^[16]。1964年7月至1993年12月期间的SOI绘于图2的下部。

观察图2可以发现,日长年际变化序列和SOI存在着较好的对应关系。根据公式(1),计算出它们之间的相关系数,得到 $\rho = -0.457$ 。由于我们采用的资料样本点数 $n = 342$ (dof = 340),则可以从表1查出, $\alpha = 0.01$ 时的相关系数临界值为0.139。显然 $|\rho|$ 远远超过0.01置信度水平的要求,所以日长年际变化序列与南方涛动指数存在很强的统计相关性。利用上述相关关系,郑大伟在国际上首次应用天文方法预测到1990年后期发生的一次El Nino事件,并被观测证实^[17]。

四、结 束 语

本文应用Monte Carlo方法,通过大量的数值计算,估算相关系数的临界值,结果列于表1、表2和图1中。本文获得的相关系数临界值表不仅容量大,而且稳定性比较好,基本能够满足我们工作中相关检验的需要,同时也可供有关人员参考使用。

美国宇航局哥达德飞行中心Dr. Benjamin Fong Chao和华东师范大学王玲玲教授与作者进行了有益的讨论,特此致谢。

参 考 文 献

- [1] 黄志宏、方积乾,数理统计方法,人民卫生出版社,1988。
- [2] 王玲玲、周纪芾,常用统计方法,华东师范大学出版社,1994。
- [3] 丁月蓉、郑大伟,天文数据的处理方法,南京大学出版社,1990。
- [4] Bendat, J. S. and Piersol, A. G., Random data: analysis and measurement procedures, Wiley-Interscience Press, New York, 1971.
- [5] Otnes, R. K. and Enochson, L. D., Applied time series analysis, Wiley-Interscience Press, New York, 1978.
- [6] Bendat, J. S., and Piersol, A. G., Engineering applications of correlation and spectral analysis, Wiley-Interscience Press, New York, 1971.
- [7] 项静恬等,动态和静态数据处理,气象出版社,1991。
- [8] 冯康,数值计算方法,国防工业出版社,1978。
- [9] 徐钟济,蒙特卡罗方法,上海科学技术出版社,1985。
- [10] Thompson, R. Y., Journal of the atmospheric sciences, Vol. 36, No. 10, p2020, 1979.
- [11] 中国科学院沈阳计算技术研究所,电子计算机常用算法,科学出版社,1976。
- [12] Lambeck, K., The Earth's variable rotation, Cambridge University Press, New York, 1980.
- [13] 郑大伟,天文学进展,Vol. 6, No. 4, p316, 1988。
- [14] 郑大伟、董大南,天文学报,Vol. 27, No. 4, p368, 1986。
- [15] Rasmusson, E. M., Amer. Scientist, Vol. 73, p168, 1985.
- [16] Chen, W. Y., Mon. Weather Rev., Vol. 110, p800, 1982.
- [17] Zheng, D. W., et al., Nature, Vol. 348, p119, 1990.

A NEW CALCULATION METHOD OF CORRELATION COEFFICIENT TEST TABLE

Zhou Yonghong Zheng Dawei

(Shanghai Observatory, Chinese Academy of Sciences, Shanghai 200030)

Abstract

The Monte Carlo method and large quantity of numerical computation were employed to acquire a threshold value table of correlation coefficients with high capacity and good stability. The table can be used for reference.

Key words correlation coefficient — Monte Carlo method