時間方向に独立でない気象データの自由度を求める簡便かつ定量的な方法について

松　山　　洋*・片　境　泰　聡*

1．はじめに

気象データでは，一般に時間方向に隣り合うデータどうしに何らかの関連があるため，自由度（独立なデータ数）がデータ数よりも少なくなることは広く知られている（例えば，松山・谷本 2008; 伊藤・見延 2010）．一方，一般的な統計学の教科書では，各々のデータが独立であることを前提として，様々な検定手法が紹介されている．そのため，気象データを用いて統計的有意性について調べる際，一般的な統計学の教科書に従って各種の検定を行なうと，自由度を大きく見積もってしまい，誤った判断を下す危険性がある．

この点に関して，松山・谷本（2008）では，対象とする時間スケールとデータの長さから自由度を決めることが提案されている．しかしながら，具体的にどうすればよいかはケースバイケースであり，手法が客観的ではない．また，伊藤・見延（2010）では一次の自己回帰モデルを用いて自由度を推定する標準的な手法が紹介されている（第 2 節参照）．しかしながら，サンプル数がいくつあればこの方法が適用できるかについては述べられていない．

本稿では，伊藤・見延（2010）で紹介されている標準的な手法のサンプル数がいくつあればよいかについて検討した．この手法では近似的に有効標本数（第 2 節）を求めることになるが，実際には，有効標本数を厳密的に求めることもできる（第 3 節）．そのため，厳密解と近似解を比較し，近似式（第 2 節の式(7)）の相対誤差と，サンプル数およびラグ 1 の自己相関係数との関係についても言及した．

* 首都大学東京 都市環境科学研究科.

—2012年 7 月11日受領—
—2012年 9 月24日受理—

2．伊藤・見延（2010）で紹介されている，自由度を推定する標準的な手法

伊藤・見延（2010, p.59–60）では，実効的に独立な標本間の時間（有効無相関時間とも言う）T_eで，データのサンプル数 N を割ることによって，有効標本数 N_eが得られると述べられている（式(1)）．この場合，例えば相関係数の検定では，有効自由度は N_e-2になる．

$$N_e=\frac{N}{T_e} \tag{1}$$

さらに伊藤・見延（2010）では，平均，分散，共分散（相関および回帰）に関する有効無相関時間を求める式がそれぞれ示されている．本稿では伊藤・見延（2010）の付録（p.225–226）との関連上，平均に関するものだけを取り扱う．ここで，平均に関する無相関時間 T_eは以下の式(2)で求められる．

$$T_e=\sum_{n=-N}^{N}\left(1-\frac{|n|}{N}\right)R_{xx}(n)=1+2\sum_{n=1}^{N}\left(1-\frac{n}{N}\right)R_{xx}(n) \tag{2}$$

ここで n はラグ，Nは標本の数（時系列の長さ），$R_{xx}(n)$はラグ n の時の自己相関係数である．そして，T_eを求める簡便な方法として，大気・海洋の解析でよく利用される一次の自己回帰モデル（式(3)）を考え，ラグ 1 の自己相関係数 R_1から有効無相関時間 T_eを近似的に表わす方法が紹介されている．

$$x(t)=R_1x(t-1)+\varepsilon(t) \tag{3}$$

ここで R_1はラグ 1 の自己相関係数，$\varepsilon(t)$ は白色ノイズである．式(3)が成り立つ時，ラグ n の時の自己相関係数 $R_{xx}(n)$ は以下の式(4)のように書ける．

$$R_{xx}(n)=R_1{}^n \tag{4}$$

これを式(2)に代入すると，式(2)は以下の式(5)のようになる．

$$T_e=1+2\sum_{n=1}^{N}(1-\frac{n}{N})R_1{}^n \tag{5}$$

伊藤・見延（2010, p.225-226）では，Nが十分大きい時，またはR_1が小さい時に，以下の近似式(6)を使って，式(5)が式(7)のように近似できると述べられている．これは，式(6)を式(5)に代入すると，式(5)の第2項以下は初項$2R_1$，公比R_1の等比級数になるためである．$|R_1|<1$の時にこの等比級数は収束し，収束値は$2R_1/(1-R_1)$になる．そのため，式(5)は式(7)のように近似できるのである（Trenberth 1984）．

$$(1-\frac{n}{N})R_1{}^n \fallingdotseq R_1{}^n \tag{6}$$

$$\tilde{T}_e=\frac{1+R_1}{1-R_1} \tag{7}$$

なお，式(7)および以下の記述では，近似であることを強調するために，近似式で求めた平均に関する無相関時間を$\tilde{T}_e$，$\tilde{T}_e$を用いて式(1)から求めた有効標本数を$\tilde{N}_e$と，それぞれ表わすことにする．

式(7)が成り立つための条件は，「Nが十分大きい時，またはR_1が小さい時」である．それでは，これらは具体的にはどのような値になるのであろうか？

3．厳密解と近似解の比較

実は，式(5)には厳密解がある．ただし，この場合の厳密解とは，対象とする時系列データが一次の自己回帰モデル（式(3)）で表わせるという前提のもと，式(6)の近似を用いない厳密解という意味である．同様に，以下では式(6)の近似を用いるという意味で，近似解という用語も用いる．ここではまず，NとR_1の関係について考察する前に，厳密解と近似解の比較を行なっておきたい．

森口ほか（1987，p.1）による以下の公式(8)を用いると，式(5)は以下の式(9)のように書ける．

$$\sum_{k=0}^{n}(a+kd)r^k=\frac{a-(a+nd)r^{(n+1)}}{1-r}+\frac{dr(1-r^n)}{(1-r)^2}\quad(r\neq1) \tag{8}$$

$$T_e=\frac{1+R_1}{1-R_1}+\frac{2R_1(R_1{}^N-1)}{N(1-R_1)^2} \tag{9}$$

式(9)は，平均に関する有効無相関時間の厳密解である．ここで，厳密解T_e（式(9)）と近似解$\tilde{T}_e$（式(7)）の関係について検討すると，式(9)においてNが無限大になる時第2項はゼロになり，式(9)は式(7)に等しくなる．また，$R_1<1$であることから，式(9)の右辺第2項は必ず負になり，T_eは必ず$\tilde{T}_e$よりも小さくなる．すなわち，T_eおよび$\tilde{T}_e$を式(1)に代入することによって，それぞれ求められるN_eと$\tilde{N}_e$が自然数であることを考慮すると，N_eは$\tilde{N}_e$よりも常に大きくなるかまたは等しくなる．実際，後述する第1表の範囲内でNとR_1を変化させた時，$N_e-\tilde{N}_e$の値は1または0になった．そのため，NとR_1の値に関わらず式(7)の近似を使うと，有効標本数は控えめに推定される．

近似解より得られる$\tilde{N}_e$を厳密解から得られるN_eと比較した時の相対誤差は，以下の式(10)で評価される．

$$\frac{N_e-\tilde{N}_e}{N_e}=\frac{2R_1(1-R_1{}^N)}{N(1-R_1{}^2)} \tag{10}$$

ここで，相対誤差（式(10)の左辺）を10%，5%，1%とした時のNとR_1の関係を第1表に示す．この表は，R_1を0.01～0.99の範囲で変化させ，第1表が埋まるようにNの範囲を変化させた結果得られたものである．第1表より，相対誤差が小さくなるほど，同じR_1に対して必要となるNの数が大きくなることが分かる．相対誤差10%から5%になる時，Nの数は

第1表　式(10)により，相対誤差10%，5%，1%で$\tilde{N}_e$を求める際に必要なNとR_1との関係．

R_1	N		
	10%	5%	1%
0.1	2	5	21
0.2	5	9	42
0.3	7	14	66
0.4	10	20	96
0.5	14	27	134
0.6	19	38	188
0.7	28	55	275
0.8	45	89	445
0.9	95	190	948

おおむね2倍となり，相対誤差5％から1％になる時には，N の数はおおむね5倍となっている．また，同じ相対誤差に対して，R_1が大きくなると N の数も大きくなるが，特に R_1の値が大きいところで N の増加率が大きくなっている．この傾向は，相対誤差10％よりも5％，1％の時に，より顕著になる．

結局，式(7)が成り立つための条件「N が十分大きい時，または R_1が小さい時」とは，第1表を見て各人が主観的に決めることになる．例えば，相対誤差10％で近似することにし，R_1が0.5の時に式(7)，式(1)によって$\tilde{N}_e$を決めるのに必要な N の数は14ということになる．

4．まとめ

本稿では，式(1)を用いて有効標本数を求めるため，式(7)による近似が成り立つ条件について検討した．平均に関する無相関時間の厳密解 T_e（式(9)）と近似解 $\tilde{T}_e$（式(7)）を比較すると，厳密解から得られる有効標本数 N_eは近似解から得られる$\tilde{N}_e$よりも常に大きくなるかまたは等しくなり，$N_e-\tilde{N}_e$の値は1または0になる．つまり，式(7)の近似を使うと有効標本数は控えめに推定されるため，現実的には式(7)，式(1)を用いて有効標本数を求めるのでよい．

本稿のオリジナリティは以上の点に尽きるが，このことは「N が十分大きい時，または R_1が小さい場合に式(7)が成り立つ」という定性的な情報ではなく，より定量的かつ有益な情報を与えると，筆者たちは信じる．

なお，本稿で述べたことは，大気・海洋の解析でよく利用される一次の自己回帰モデル（式(3)）で表現される現象に限られることに注意されたい．このような現象の自己相関係数を図化すると釣鐘型の分布になるが，例えば南方振動指数の場合はこのようにはならず，自己相関係数が正負両方の領域にまたがる（例えば，松山・谷本 2008）．このような場合でも式(7)，式(1)が適用可能かどうか，今後検討する必要がある．

謝　辞

草稿に対して，谷本陽一さん（北海道大学大学院環境科学院）からコメントをいただきました．また，査読者からいただいたコメントによって，本稿は大幅に改善されました．厚く御礼申し上げます．

参　考　文　献

伊藤久徳，見延庄士郎，2010: 気象学と海洋物理学で用いられるデータ解析法．気象研究ノート，(221)，253pp.

松山　洋，谷本陽一，2008: UNIX/Windows/Macintoshを使った実践 気候データ解析 第二版．古今書院，126 pp.

森口繁一，宇田川銈久，一松　信，1987: 岩波数学公式 II 級数・フーリエ解析．岩波書店，340pp.

Trenberth, K. E. 1984: Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations. Mon. Wea. Rev., **112**, 2359–2368.

A Simple but Quantitative Method to Objectively Determine the Degree of Freedom, Applicable to Serially-Correlated Meteorological Data

Hiroshi MATSUYAMA* and Hiroaki KATASAKAI**

* (*Corresponding author*) *Graduate School of Urban Environmental Sciences, Tokyo Metropolitan University, 1-1, Minami-Osawa, Hachioji, Tokyo 192-0397, Japan.*

** *Graduate School of Urban Environmental Sciences, Tokyo Metropolitan University.*

(Received 11 July 2012; Accepted 24 September 2012)