4-4　時系列データ解析

東京大学 数理・情報教育研究センター

2020年5月11日

概要

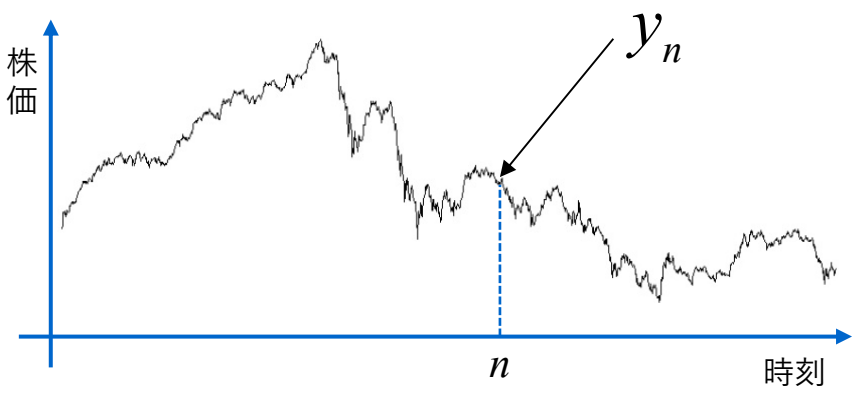
- 本節では，まず時系列とは何か，時系列データ解析の目的は何かなど時系列データ解析の概略について学びます.
- 次に，時系列データがもつトレンド，周期性，季節性，ノイズについてその意味を学ぶとともに、移動平均，階差などによる情報抽出の方法とスペクトや相関関数による特徴可視化を学びます.
- さらに，時系列モデルを用いた予測の方法と変数変換についても簡単に説明します.

本教材の目次

時系列とは

身の回りの様々なデータ

- 気温，気圧
- 株価，為替レート
- GDP, 消費者物価指数
- 血圧，脈拍，脳波
- 地震の波動

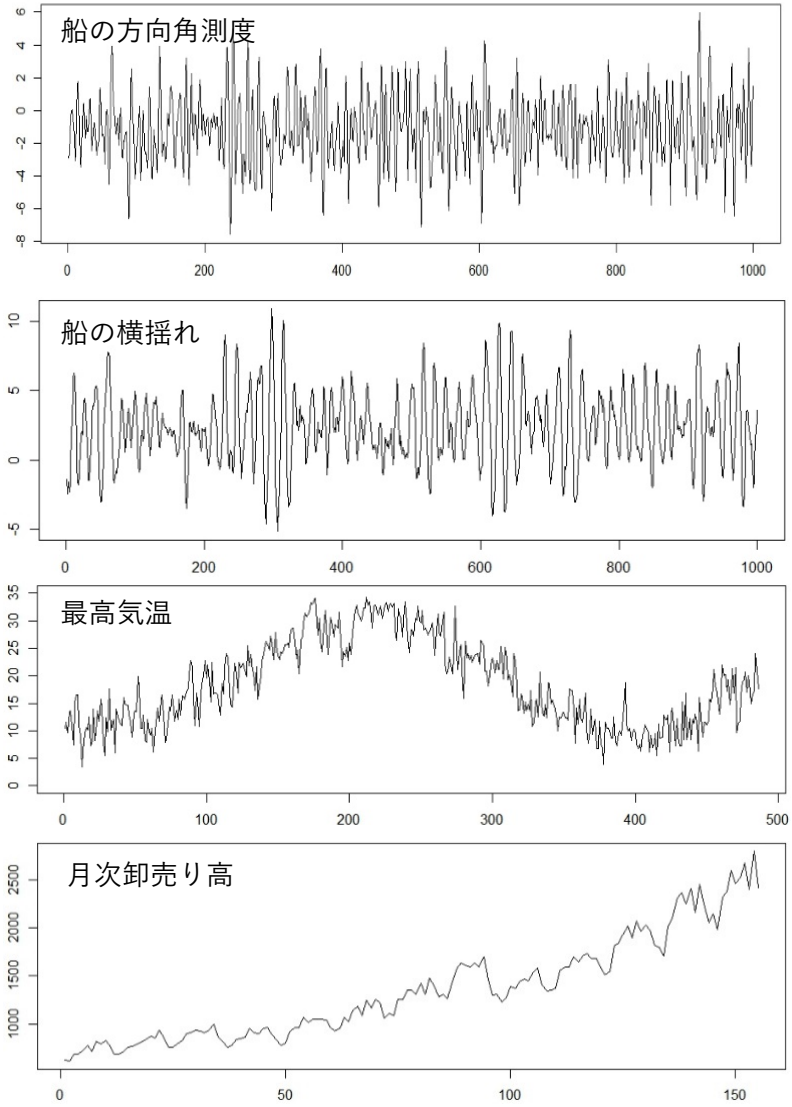


などは時間とともに変動しています．

このように時間とともに変動している現象の記録が時系列です．以下では時系列データを $y_1,\ldots,y_N$ と表します．

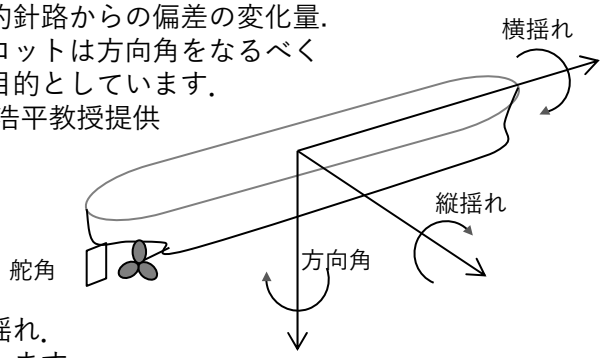
- 時系列を図示するとき横軸は時刻 n，縦軸は時系列の値 y_n です．
- N はデータ数（時系列の長さともいいます）
- データの測定間隔（Δt ）は，年，月，日，時間，秒，1/100秒などデータによって様々です．

いろいろな時系列の例



船の方向角速度

航行中の船舶の目的針路からの偏差の変化量.
船舶のオートパイロットは方向角をなるべく
小さくすることを目的としています.
東京海洋大学 大津浩平教授提供
N=1000, ΔT=1秒



船の横揺れ

船舶の左右方向の揺れ.
乗り心地にも影響します.
10秒程度のゆっくりした変動が見られます.
N=1000, ΔT=1秒

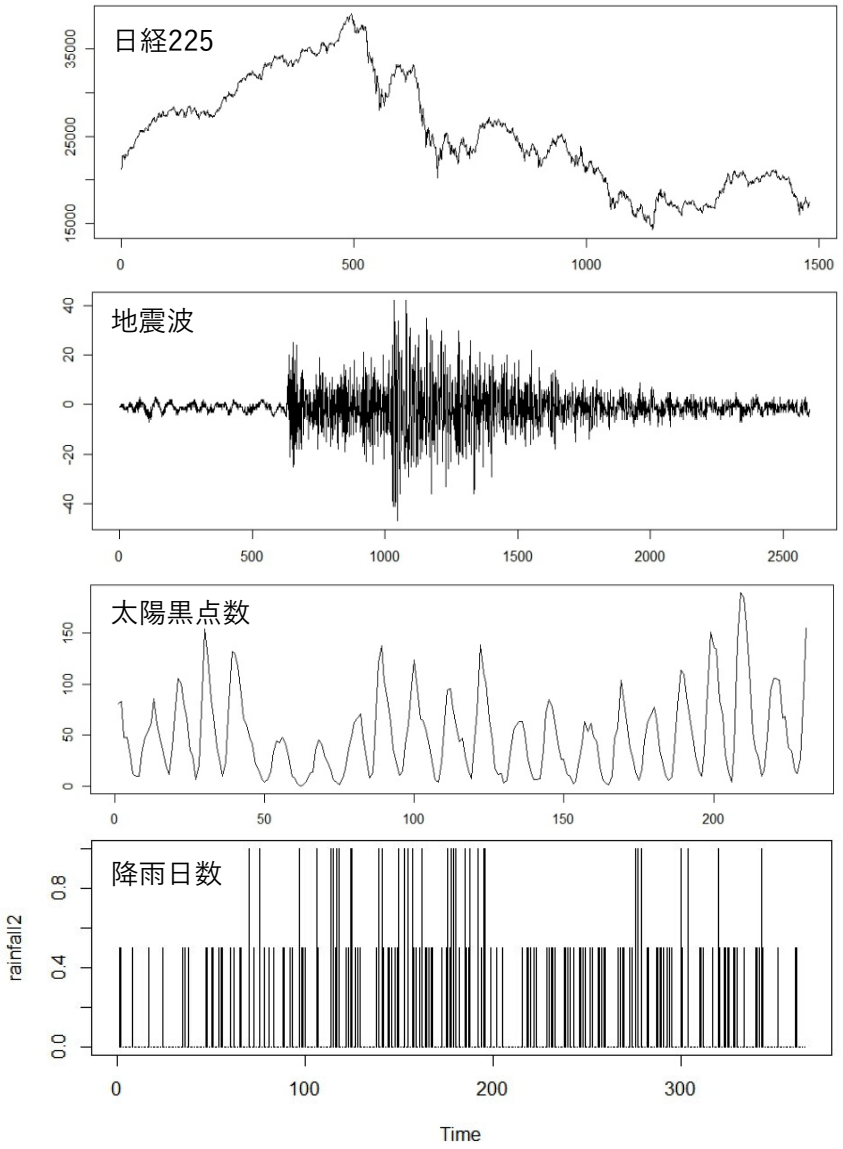
太陽黒点数

東京の1日の最高気温を観測したもの.
１年周期と短期的変動の合成のように見えます.
気象庁公表データ
N=500, ΔT=1日

月次卸売り高

あるハードウェアの毎月の卸売り高.
増加傾向と1年の周期的変動が見えます.
平均の増大とともに変動幅も段々大きくなっています.
合衆国BSL公表データ
N=155, ΔT=1月

いろいろな時系列の例（２）



日経225

日経225平均株価.
ゆっくり上下する傾向変動成分とその周りの短期的変動の合成のように見えます．短期的変動の大きさは場所によって異なっています.
N=1440, ΔT=1日

地震波

平均の変化は見られないが変動の幅はP波とS波の到着によって2か所で増大し，その後だんだん小さくなっています.
北海道大学 高波鐵夫氏提供
N=2600, ΔT=0.02秒

太陽黒点数

Wolfer太陽黒点数データと呼ばれるデータ．11年程度の周期が見られますが，波形は上下非対称で前後も非対称．変動が小さな時は周期が長くなる傾向がみられます.
N=231, ΔT=1年

降雨日数

東京で2年間の1月1日～12月31日に雨が降った日数.
0, 1, 2の離散値データとなることが特徴．季節によって降雨日数の変化が見られます.
気象庁公表データ.
N=365, ΔT=1日

多変量時系列

互いに関連する複数の時系列が同時に記録されたものを多変量時系列と呼びます．多変量時系列は例えば次のような状況で得られます．

・国のGDP, 消費者物価指数などの経済指標を同時に測定した場合
・特定の観測点で気圧，気温，風速，雨量などを同時観測した場合
・いくつかの異なる観測点で地震の波形を同時記録した場合

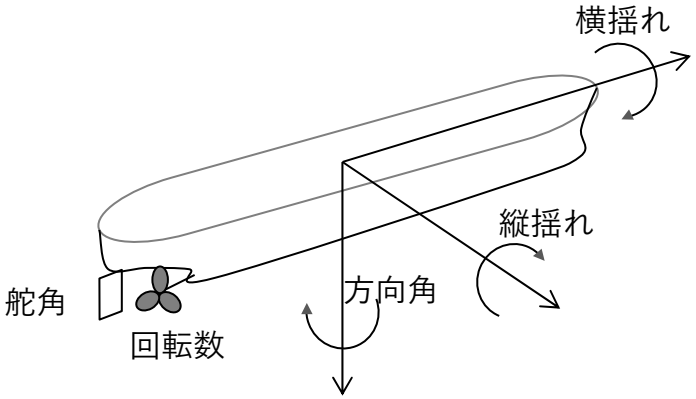
多変量時系列は通常同時に観測された時系列を縦に並べてベクトルとして考えます．

$$y_n = \begin{bmatrix} \text{気圧} \\ \text{気温} \\ \vdots \\ \text{雨量} \end{bmatrix} \Bigg\} \text{時刻 } n \text{ の時系列を} m \text{ 個縦にならべたもの．}$$

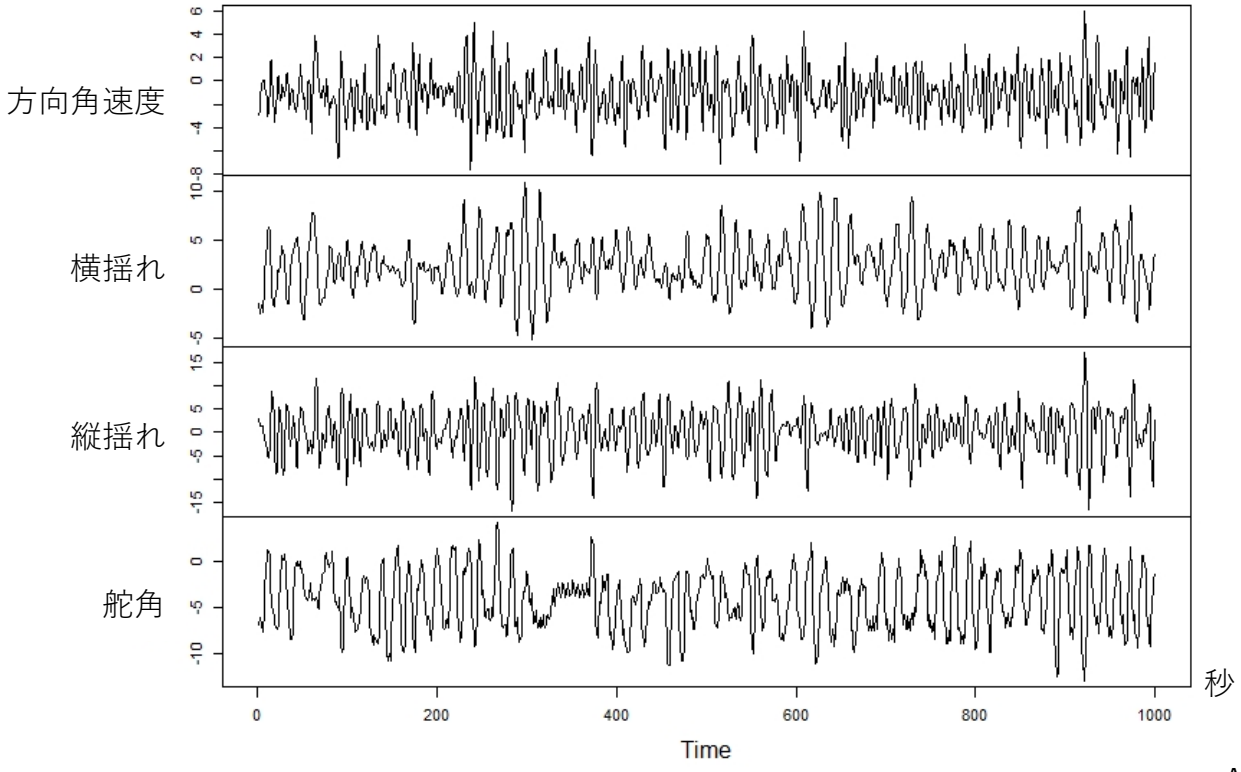
上から j 番目の成分を $y_n(j)$ と表します．m が具体的にわかっている場合には m 変量時系列と呼ぶこともあります．

多変量時系列：船舶データの例

航行中の船の揺れや舵角を同時に記録すると多変量時系列が得られます.



船の多変量データ



多変量データから船の運動特性や舵角に対する船体の応答などがわかるだけでなく，自動操舵システム（オートパイロット）の設計を行うこともできます.

時系列解析の目的

時系列解析では次のような問題を考えます.

1．可視化（時系列の特徴を捉える）
- ・データのプロット
- ・周期性をみる
- ・時間的な相関をみる

2．情報抽出（時系列から情報を取り出す）
- ・トレンド
- ・季節成分
- ・ノイズ

3．予測（時系列の将来を予測する）
- ・自己回帰モデル

4．意思決定，制御

＊モデルカリキュラムでは１と２だけを取り上げていますが
本教材では３までをカバーしています.

時系列から情報を取り出す

どんな情報に注目するかによって分析方法が違います

時系列の傾向を見たい → 小さな変動を無視してトレンドを推定する

【方法】・移動平均

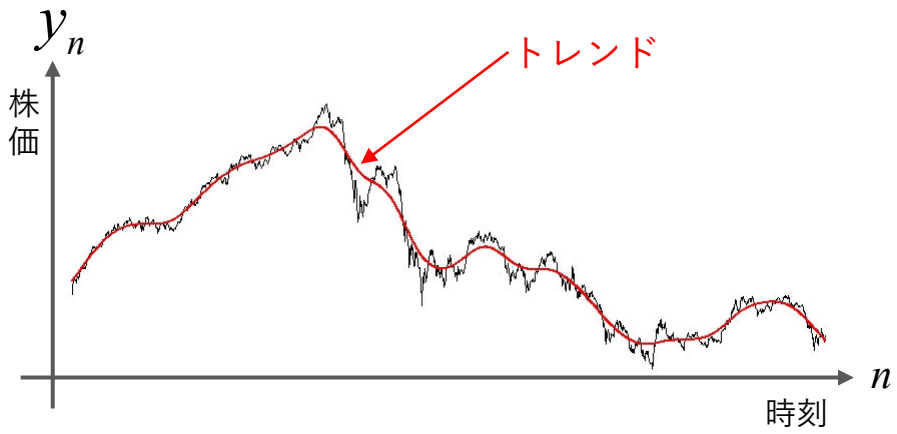
・季節調整

短期的な動きを見たい → トレンドを除去して，トレンド周りの動きを抽出する

【方法】・階差

トレンド

時系列の長期的傾向をトレンドと呼びます.



- トレンドは経済動向の分析などに使われます.
- トレンドの推定方法はいろいろあります．目的や推定方法によって，トレンドの滑らかさは変わります．本教材では
  - 移動平均
  - 季節調整

  を紹介します.

移動平均

以下の式で平滑化してゆっくりした傾向変動を取り出すための方法です.

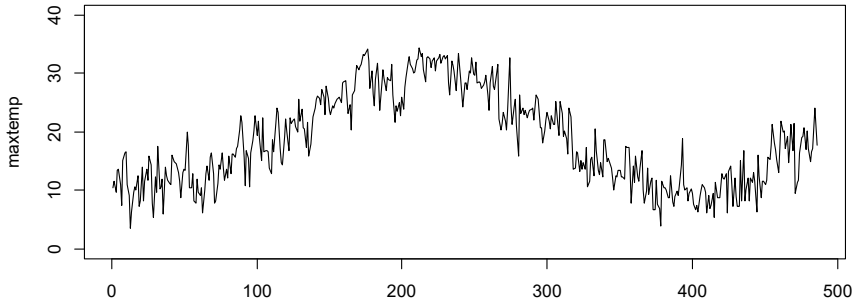
$$T_n = \frac{1}{2k+1}\left(y_{n-k} + \cdots + y_n + \cdots + y_{n+k}\right)$$

- 移動平均は y_n と前後 k 個（合計$2k+1$個）のデータの平均です
- $K = 2k+1$ は項数と呼ばれます
- 音声や画像の信号処理，金融のテクニカル分析などで利用されます

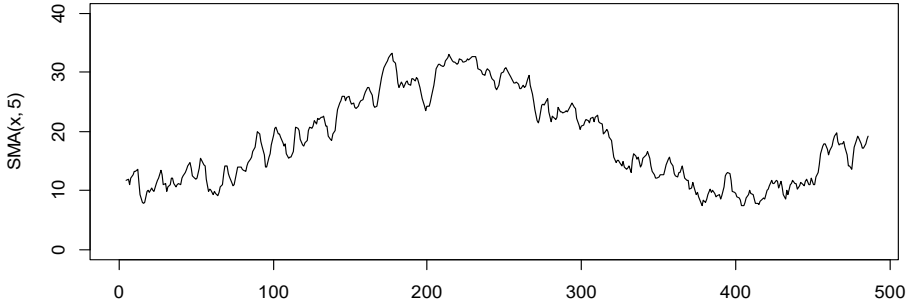
※ 移動平均を計算すると変動が小さくなります．（直線にノイズが加わっているデータの場合には）変動の分散が$1/K$ に減少します.

移動平均の計算例

この気温データの移動平均を計算します



5項移動平均



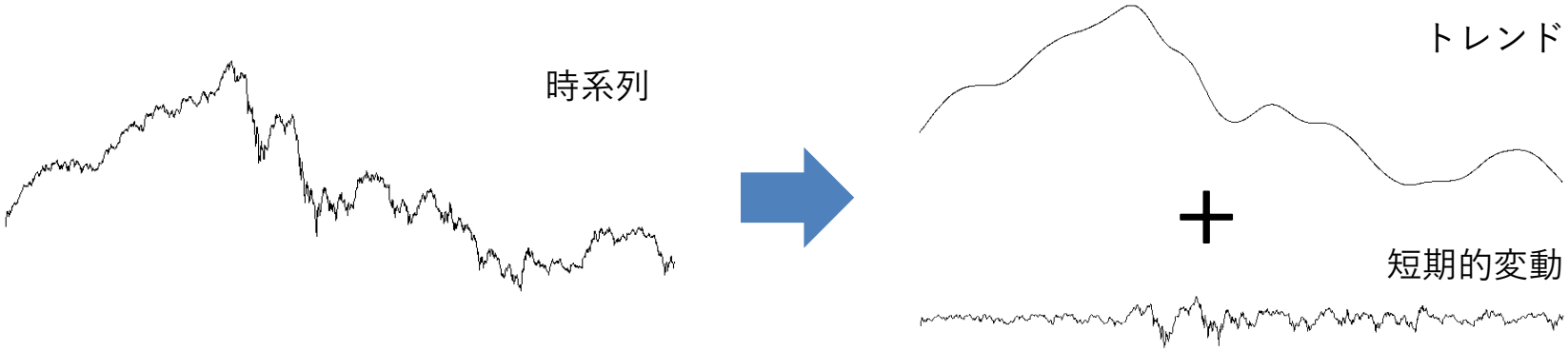
17項移動平均

29項移動平均

- 項数 K を増やすと段々滑らかになります．項数 Kの選び方が重要ですが，客観的に決めるのは難しいです．
- モデルを使うともっと綺麗なトレンドが得られ，滑らかさも自動的に決めることができます．

ノイズ

一方，時系列のうち取り出したい情報以外の不要な情報はノイズと呼ばれます．



時系列を２つの成分に分解するとき，何を「取り出したい情報」何を「ノイズ」とするかは，解析の目的によって変わります．

- 時系列の長期的傾向を見たいとき
  ⇨ トレンドが必要な情報，短期的変動はノイズ．
- 直近の動きを見たいとき
  ⇨ 短期的変動が必要な情報，トレンドはノイズ．

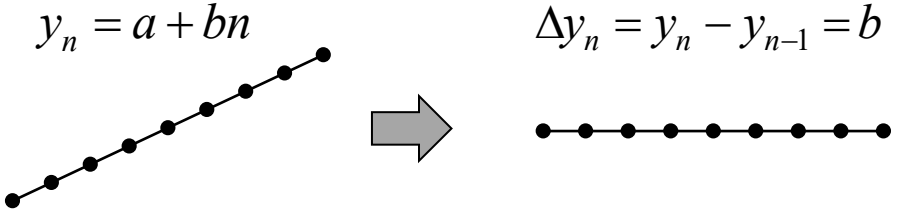
* 時系列解析では重要なノイズとして白色雑音があります．（後述）

階差

階差は時系列からゆっくりした変動を除去する簡単な方法です.

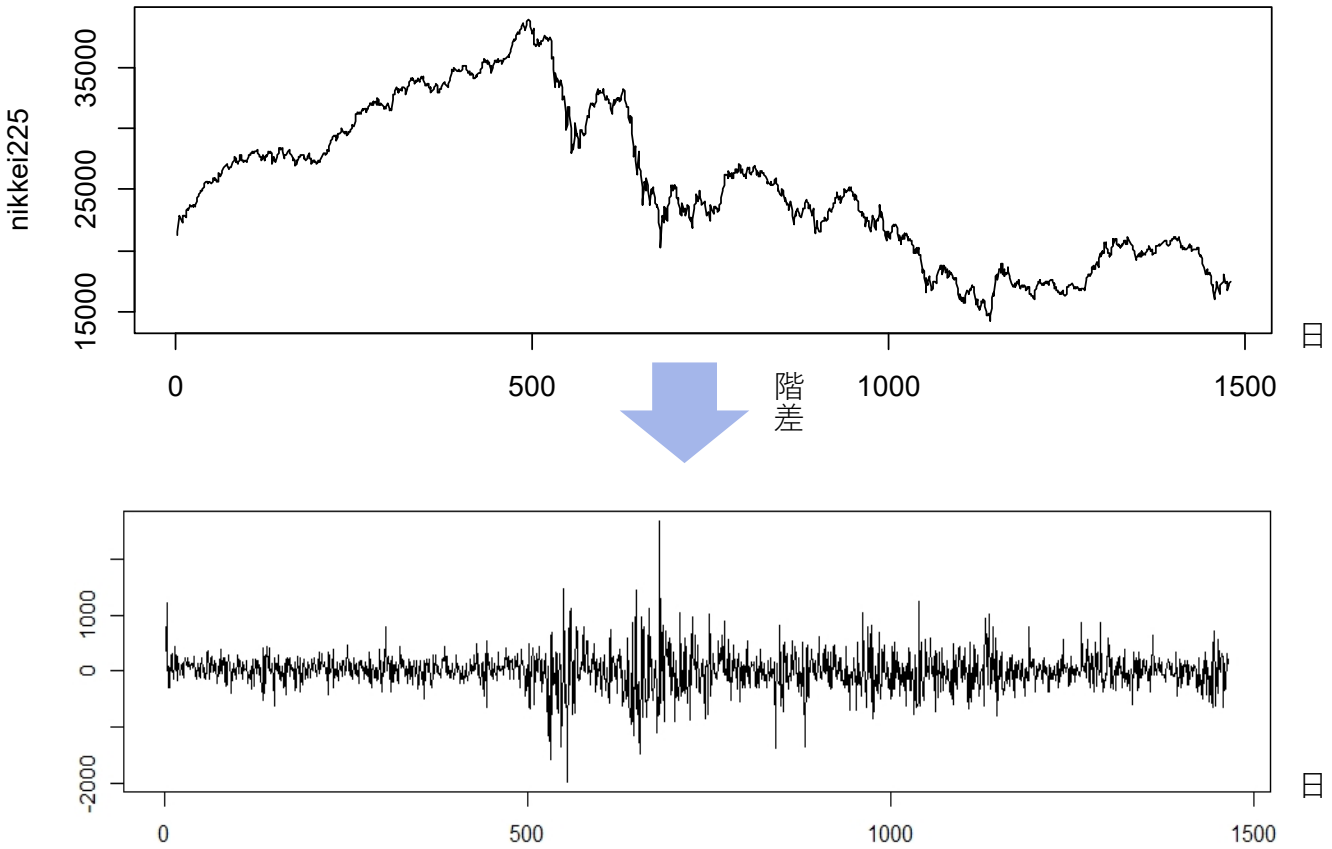
階差とは $\Delta y_n = y_n - y_{n-1}$ という操作，すなわち 1 時点前の値との差を計算することです.

- 差分とも呼ばれます
- 階差によって直線的な動きは除去できます
  （非定常な時系列は階差によって定常化できることが多い）



- 階差を2回適用すると2次曲線が除去できます（2階階差）

例：日経225平均株価データの階差



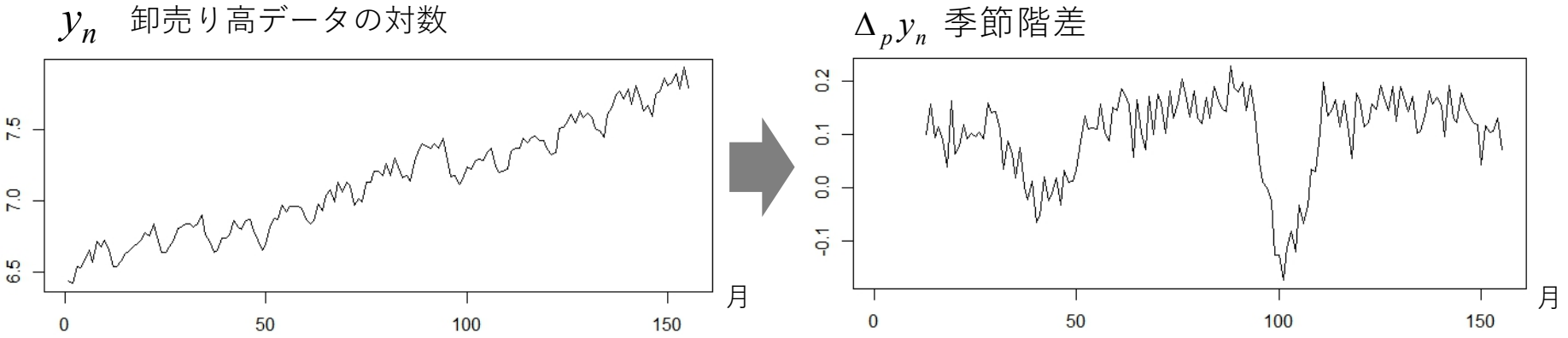
- 階差によってトレンドが除去され，短期的変動が浮き彫りになります
- 変動（ボラティリティ）が大きなところがわかります

季節階差

時系列に季節成分がある場合，周期 p の階差をとる季節階差によって季節変動をある程度除去できます．

$$\Delta_p y_n = y_n - y_{n-p}$$　　季節階差（差分）

p は季節の長さで，月次データの場合 p=12です．



季節階差によって，n=40と100付近の落ち込みが明確になります．

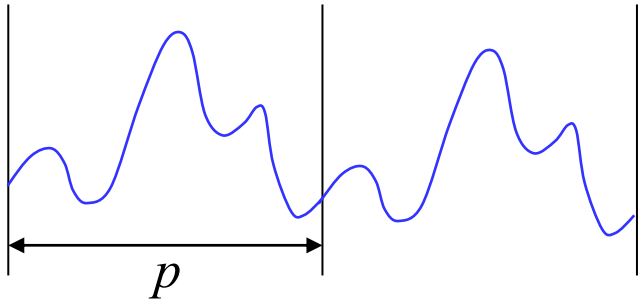
※ 経済時系列の場合は，階差と季節階差の代わりに，前期比と前年比を使うことがあります．

時系列の周期

時系列が一定の間隔で同じような変動を繰り返す成分を持つとき，
周期的変動と呼びます．

$s_n \simeq s_{n-p}$　　周期的変動

p　　周期



- 経済時系列では1年周期の変動は季節変動と呼ばれます．
- 現実の時系列では，$s_n=s_{n-1}$ が厳密に成り立つ完全な周期変動は少なく周期的変動パターンは徐々に変化します．
- 完全な周期関数の推定は簡単ですが，少しずつ変化する周期的成分の推定は少し面倒です．

時系列の特徴を可視化する方法

時系列の特徴を可視化する方法は２つあります

- スペクトル

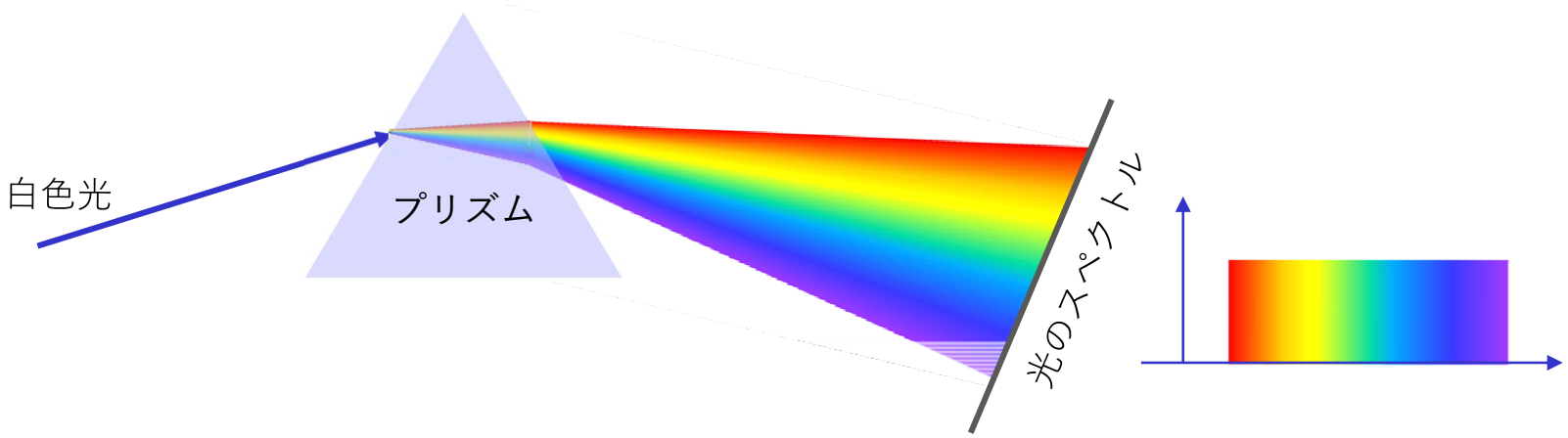
  時系列を周期関数に分解して，その大きさを見る

- 相関関数

  離れた時点の時系列との相関を見る

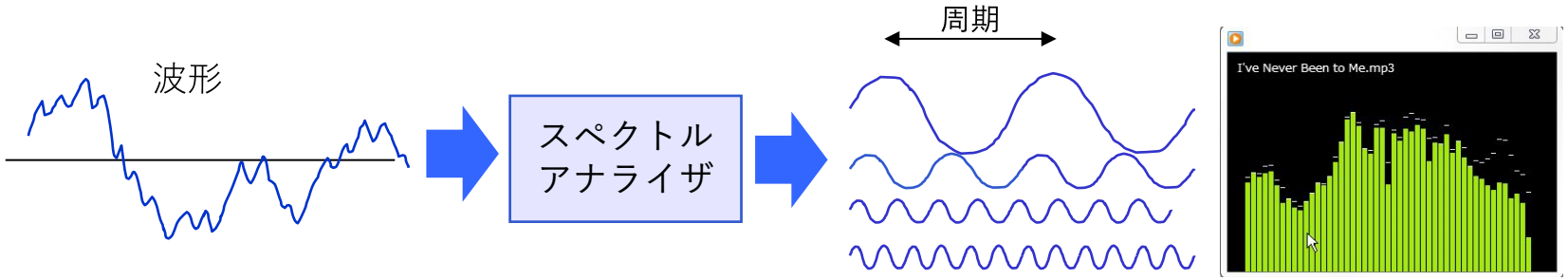
スペクトル

光をプリズムで沢山の色に分解できるように，時系列はサイン・コサインの和に分解できます．その強さを表示するとどの周期が強いかが分ります



同じようにランダムな波形も周期関数の和で表現できます。

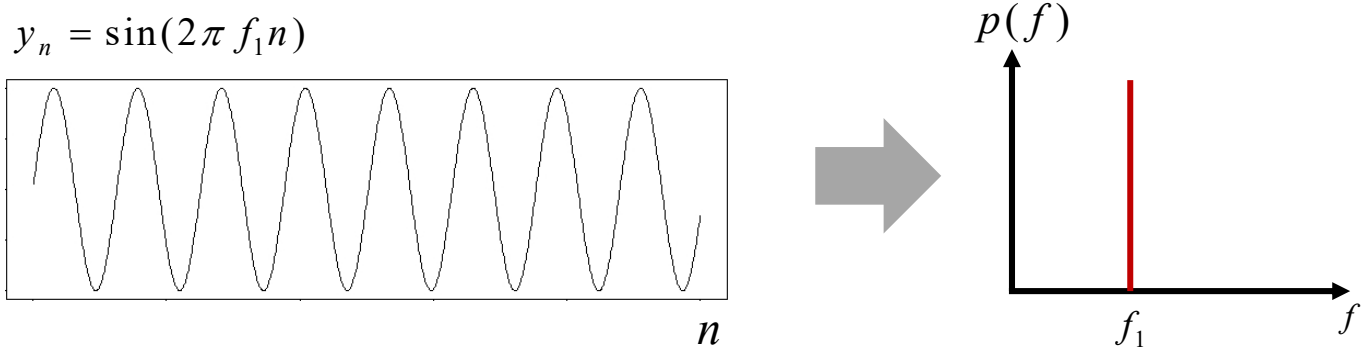
- スペクトルアナライザや音声のグラフィック・イコライザが例です



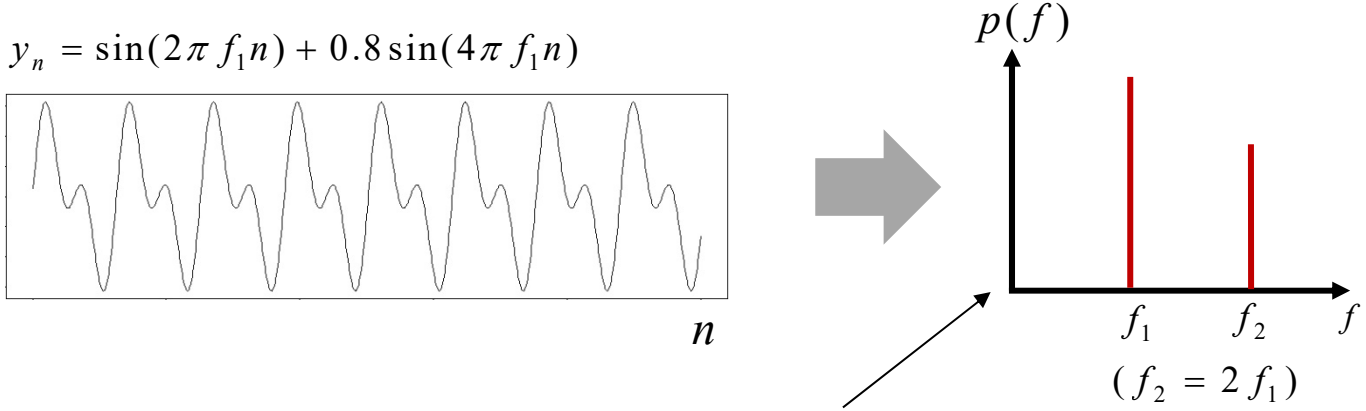
周波数=1/ 周期

周期関数の場合

- サイン・コサインのようなひとつの周期関数で表現できる場合にはスペクトルは1点に集中します.



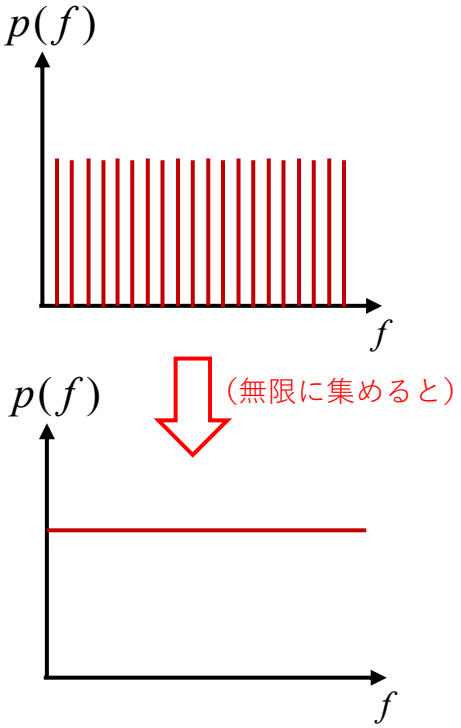
- ２つの周期関数の和の場合はスペクトルは２つになります.

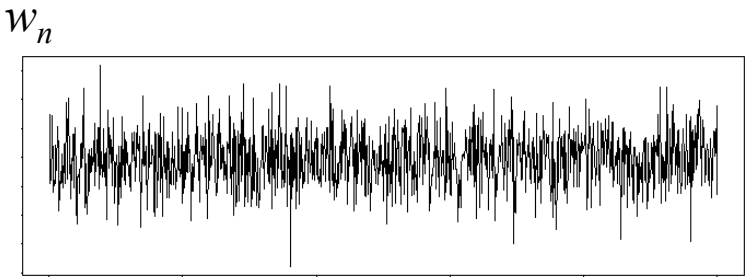


縦棒の高さは振幅の2乗に比例します.

白色雑音

- 周期関数を同じ割合で足していったらどうなるでしょうか？
- あらゆる周期の波が同じ割合で含まれているので，
  スペクトルは一定値になります．
- このようなあらゆる周期の波が均等に含まれる時系列は
  白色雑音と呼ばれ理想的なノイズと考えられています．



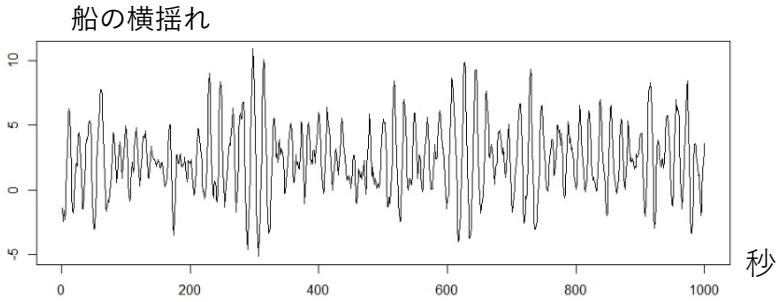


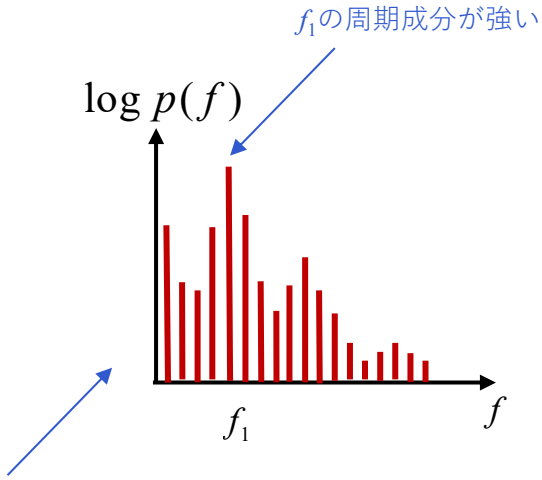
白色雑音の例

※ 白色光はいろいろな波長の光が同じ割合で混ざったものです．

現実の時系列の場合

- 現実の時系列は周期関数と白色雑音の中間で，多数の波が異なった割合で含まれています．
- したがって，たくさんの縦棒が異なった高さで並びます．

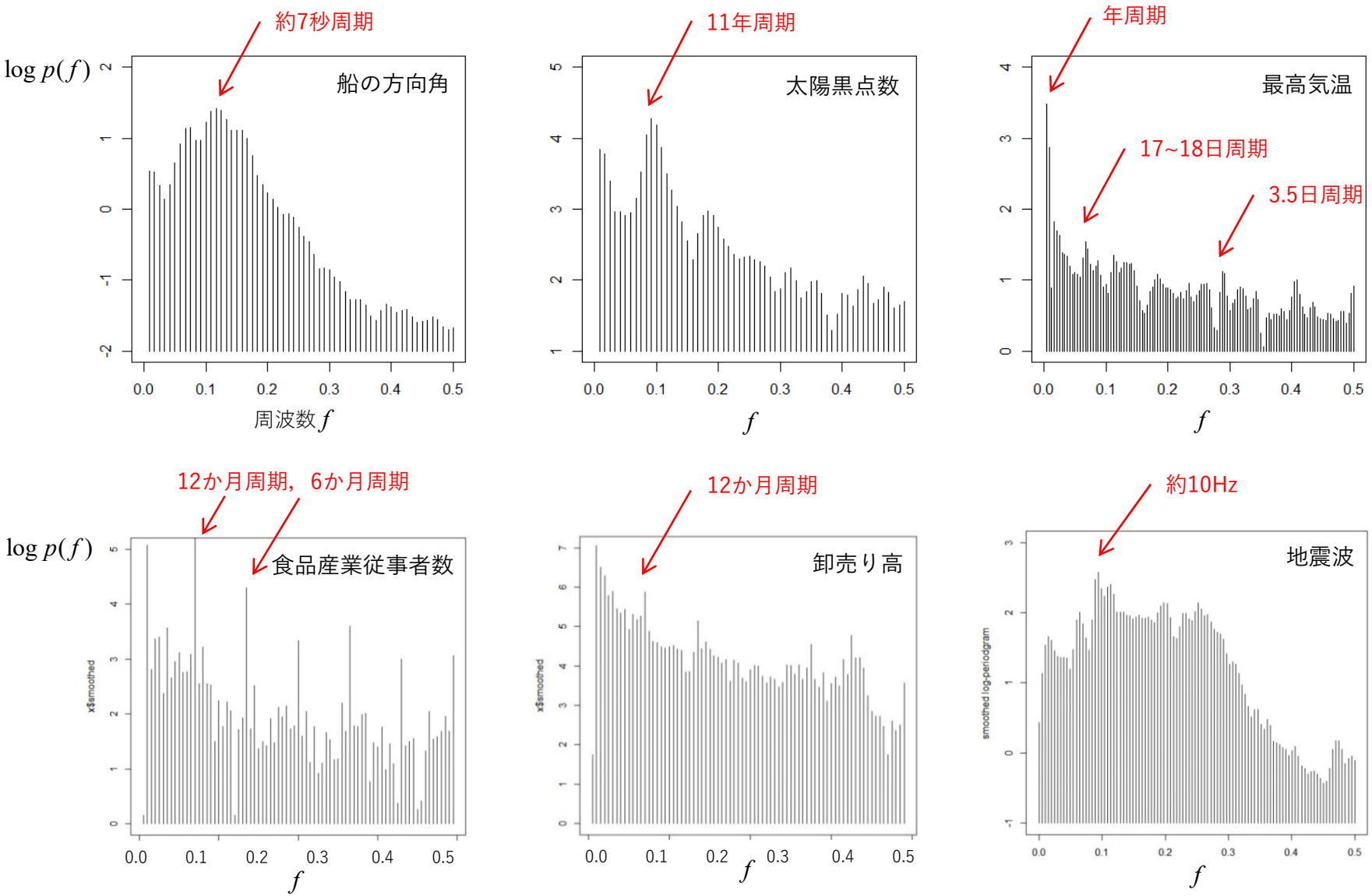




＊縦軸は対数目盛の場合が多いので注意しましょう

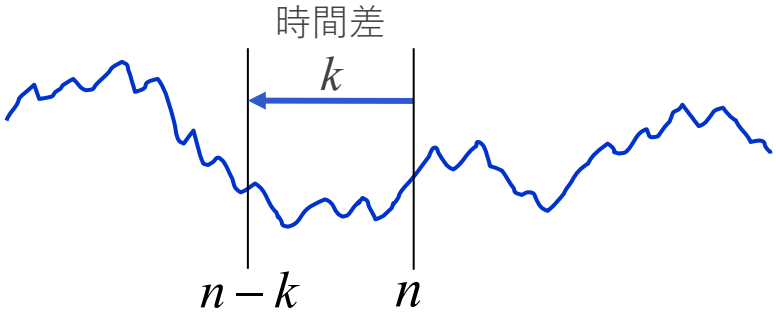
ここでスペクトルと呼んでいるのは厳密にいうとピリオドグラムですが，
ここでは厳密な使い分けはしないことにします．

実際の時系列のスペクトルの推定例



自己相関

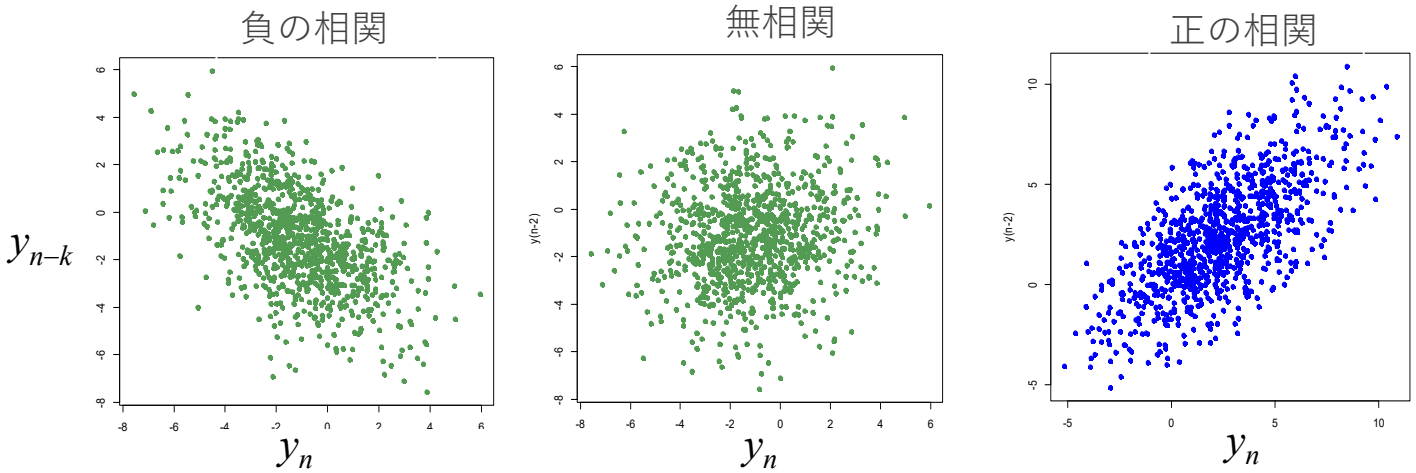
時系列のある時刻と k だけ離れた時刻との相関をみると相関の観点から時系列の特徴を捉えることができます



y_n と y_{n-k} の相関係数を $R(k)$ と書きます.

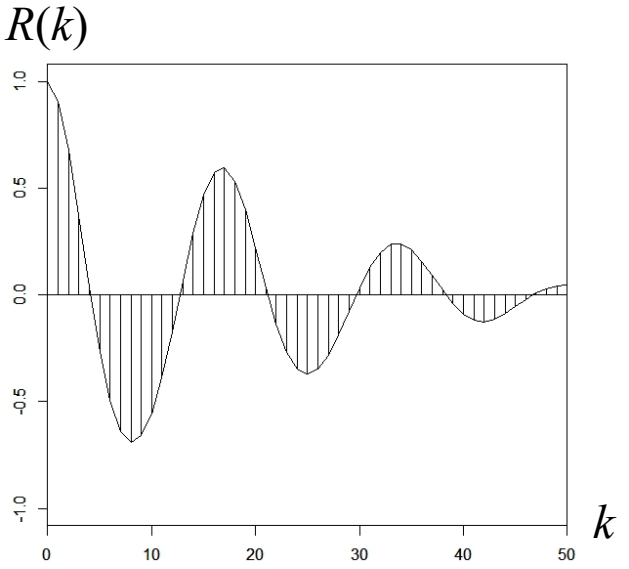
$$R(k) = \mathrm{Cor}(y_n, y_{n-k})$$

- $R(k)$ は k だけ離れた地点との相関の大きさを表します.
- y_n と y_{n-k} に正の相関があるとき $R(k) > 0$
- y_n と y_{n-k} に負の相関があるとき $R(k) < 0$



自己相関関数

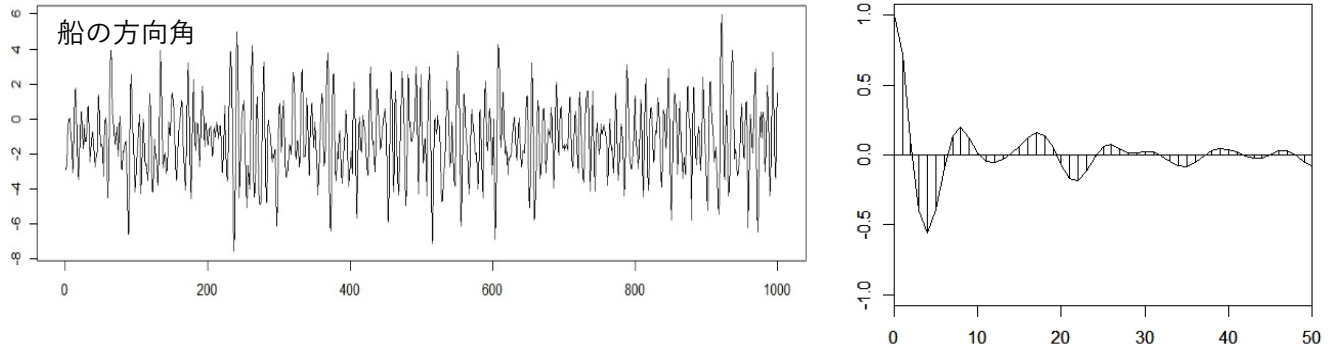
$$R(k) = \mathrm{Cor}(y_n, y_{n-k})$$



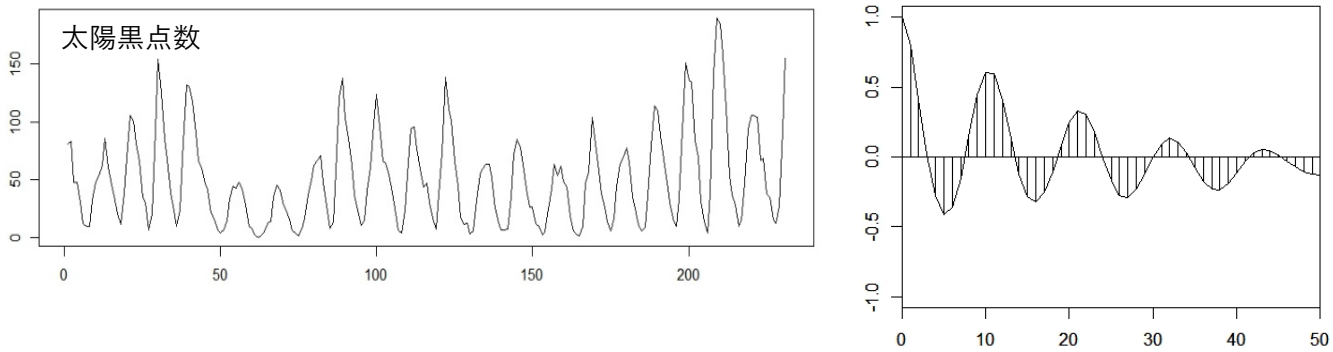
$R(k)$ を時間差 k の関数とみなしたものが自己相関関数です

- 時間差 k はラグと呼ばれます
- 周期的変動があるとその周期で $R(k)$ も上下します
- 通常，$R(k)$ は k が正または 0 の時しか図示しないのは偶関数で $R(k) = R(-k)$ が常になりたつからです

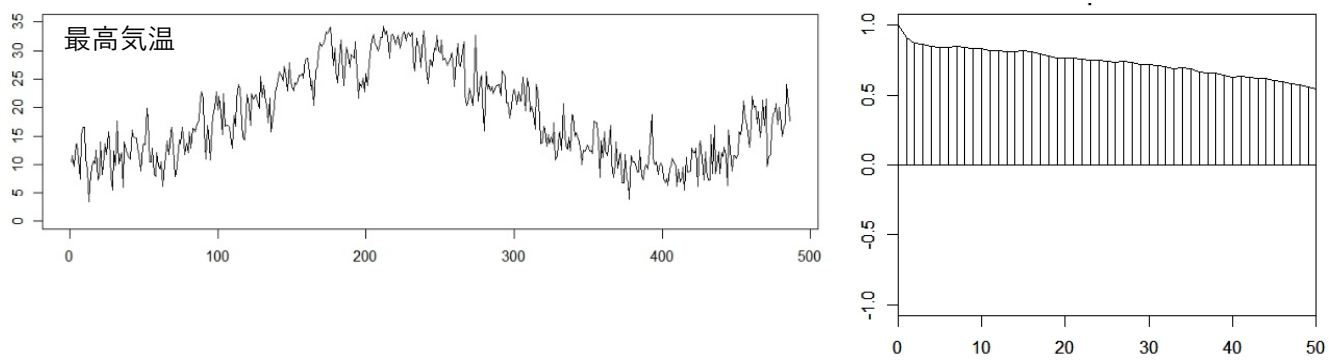
自己相関関数の例



最初の半周期だけは相関が強いが，それより先は目立った相関は見られません．



自己相関関数がいつまでも振動していて，周期性が強いことがわかります．



非常にゆっくり減衰するのは年周期のせいです．このまま分析するのは良くないことがわかります．

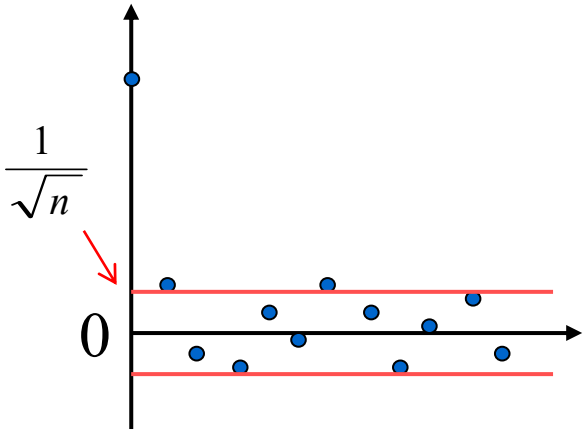
白色雑音

時間的に相関がない時系列（$R_k = 0, k \neq 0$）を白色雑音（ホワイトノイズ）と呼びます.
スペクトルのところでも出てきました.

- 理想的なノイズと考えられます
- 時系列モデルは残差が白色雑音になるように構成します.

白色雑音の自己共分散関数

時系列が白色雑音の場合，データ数がnのとき自己相関関数の推定誤差分散は$1/n$（標準偏差は$1/\sqrt{n}$）となります.

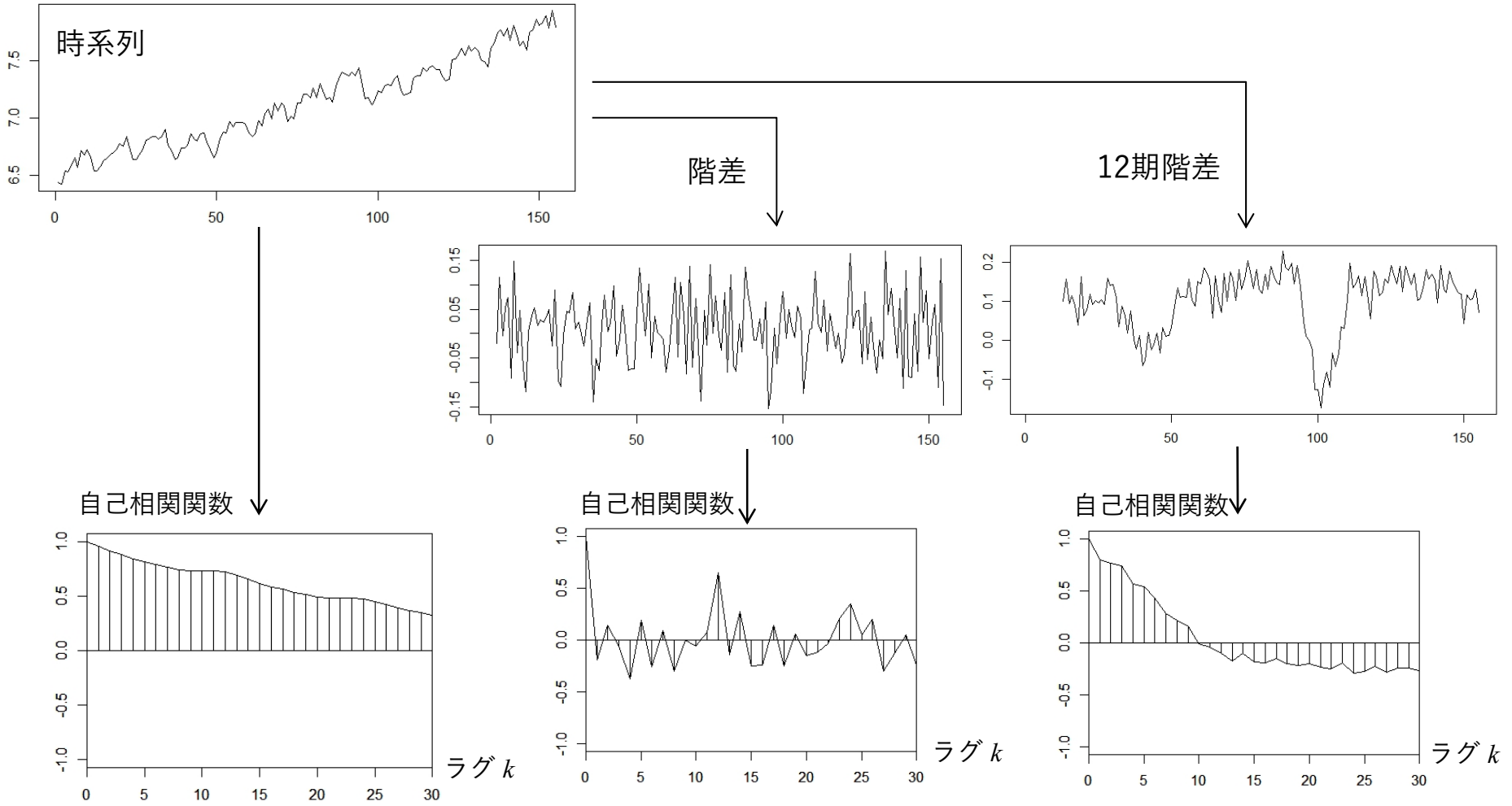


この性質は時系列が白色雑音かどうかを判断するときに使えます.

n	100	1000	10000
$\frac{1}{\sqrt{n}}$	0.1	0.03	0.01

データ数 n と推定誤差の関係

階差による相関関数の変化



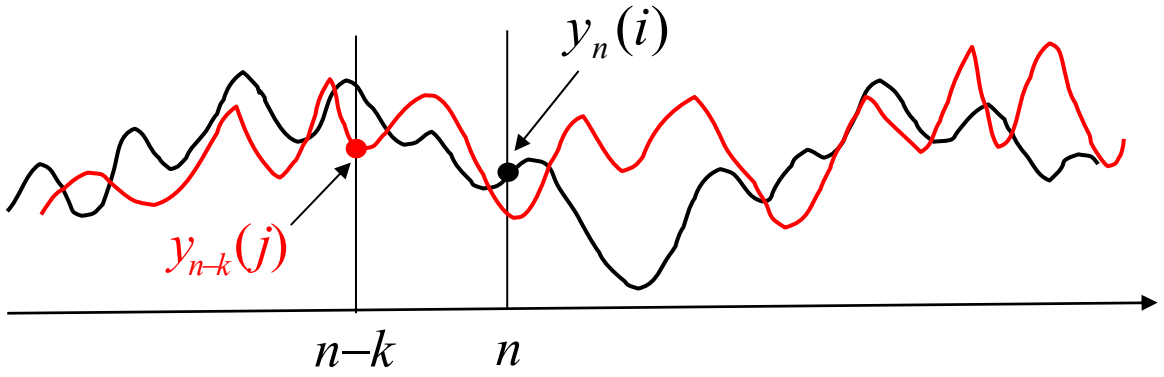
- 階差をとると，自己相関関数は著しく変化します．
- 階差によって時系列の性質が変わるからです．

相互相関

１変量の時系列と同様に，多変量時系列の場合にはi番目の時系列$y_n(i)$と時間が k だけ離れた j 番目の時系列 $y_{n-k}(j)$ の相関係数

$$R_k(i,j) = \mathrm{Cor}(y_n(i), y_{n-k}(j))$$

を考えます．これを相互相関と呼びます．



相互相関関数

m 変量時系列の場合 m×m 行列 R_k が定義できます.

$$R_k = \begin{bmatrix} R_k(1,1) & \cdots & R_k(1,m) \\ \vdots & \ddots & \vdots \\ R_k(m,1) & \cdots & R_k(m,m) \end{bmatrix}$$

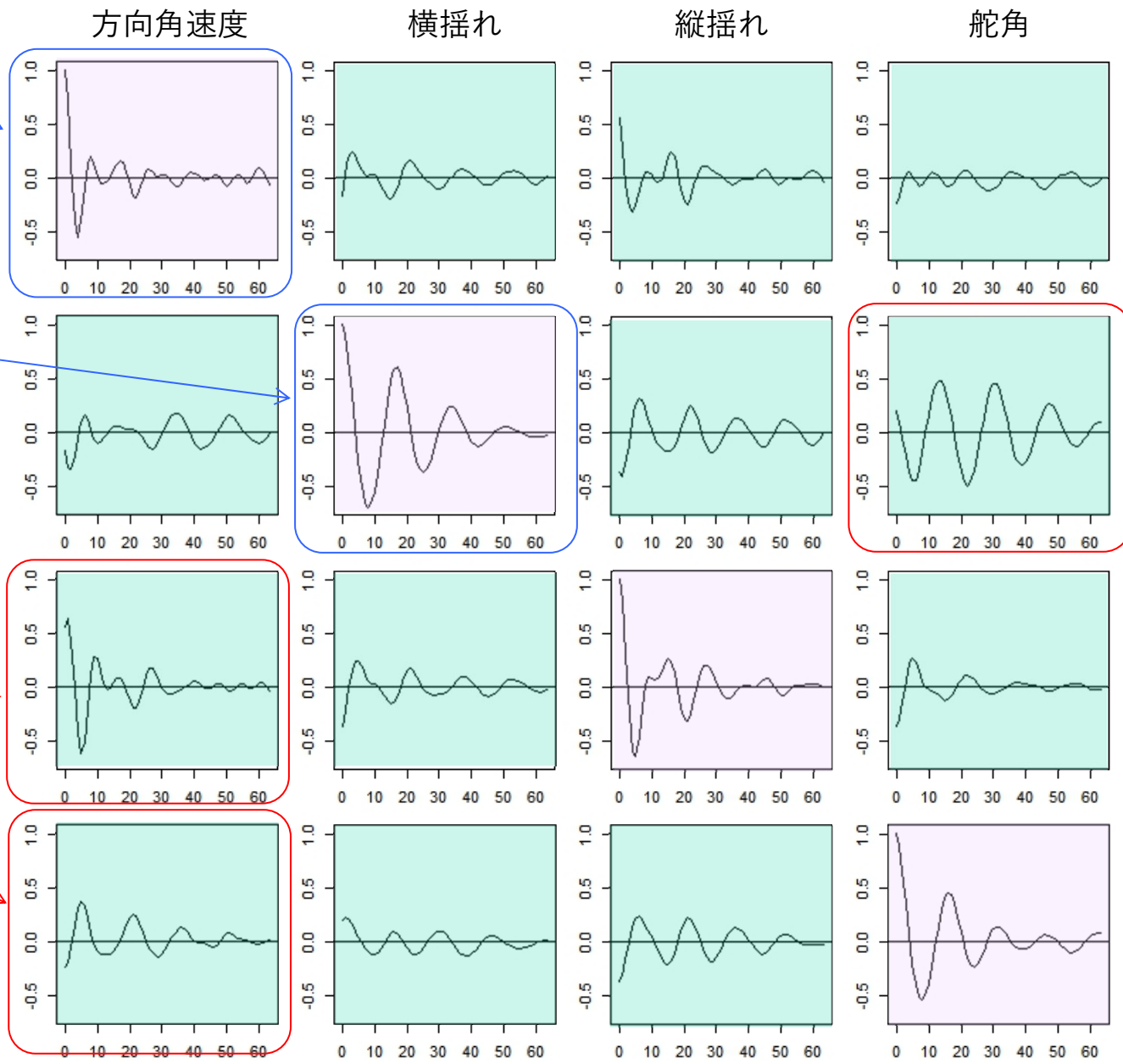
- R_k を時間差 k の関数とみなしたものが相互相関関数です.
- R_k の対角成分 $R_k(i,i)$ はi 番目の時系列 $y_n(i)$ の自己相関関数です.
- R_k の(i, j)成分 $R_k(i,j)$ は $y_n(i)$ とk時刻前の $y_{n-k}(j)$ の相関の大きさを表します. ただし, 相互相関は因果関係を示している訳ではないことに注意してください.

相互共分散関数（船舶データ）

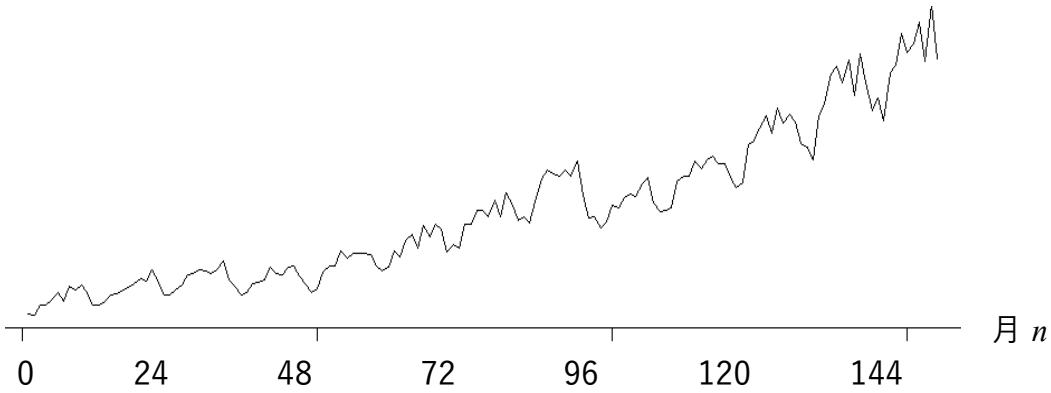
自己相関関数

相互相関関数

横軸：ラグ(秒)
縦軸：相関



季節調整とは



（月次の）経済データには，しばしば上昇や下降のトレンドと毎月同じような傾向を持つ季節成分が含まれています．

したがって，原データを見ただけでは景気の動向や販売量の増減などを的確に判断することが困難なことがあります．

季節調整とは何らかの原因で特定の周期で繰り返す成分を除去して本質的な現象を抽出することです．

季節調整の方法

季節調整では観測データ y_n をトレンド，季節成分，不規則成分（ノイズ）の3成分に分解します

$$y_n = t_n + s_n + w_n$$

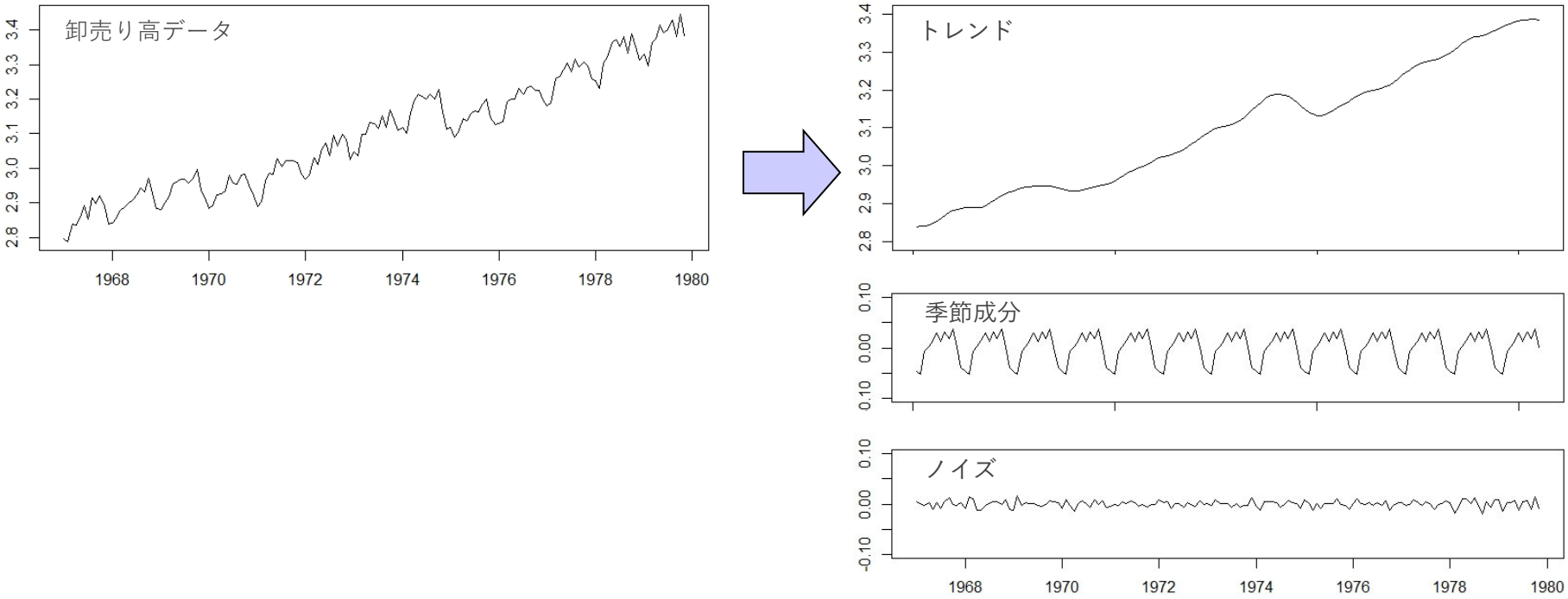
t_n　トレンド成分
s_n　季節成分
w_n　不規則成分（ノイズ）

季節成分を分離することにより，単なるトレンド推定よりトレンドの微妙な変化を捉えることができるようになります.

この方法は月次の経済時系列だけでなく，他の周期を持つデータなどへも適用できます.

気象データ（24時間），環境データ（24時間，1週間）
営業データ（1週間），4半期経済データ（４期）

卸売り高データの季節調整例



データをトレンドと季節成分とノイズに分解すると時系列の傾向がはっきり見えてきます.

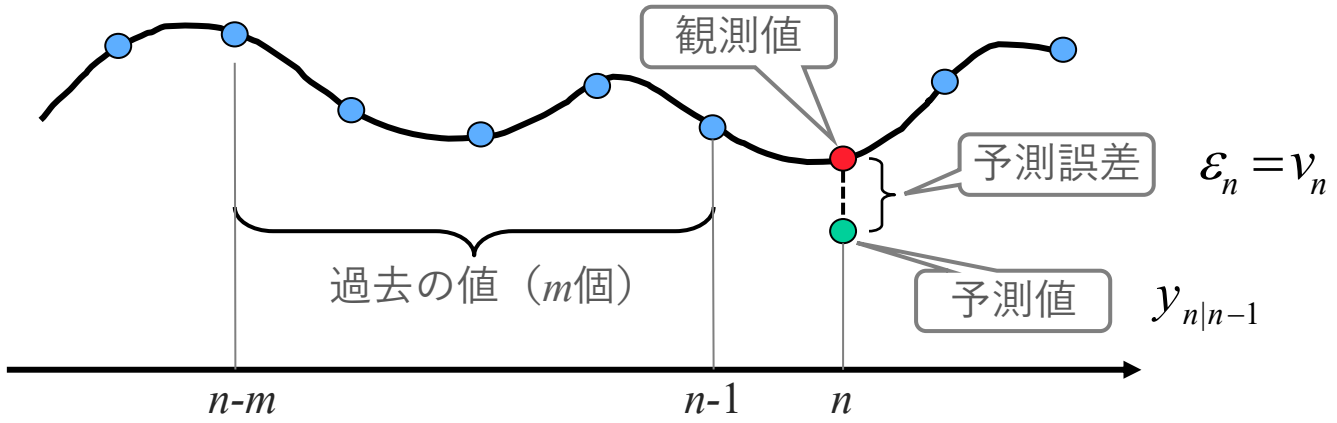
国などから公表される「季節調整済みデータ」は観測データから季節成分を除去したものです.

以上でモデルカリキュラムのスキルセットをカバーしています．以下の（発展）はスキルセットにない内容を含みます．

5. 時系列の将来を予測する：ARモデル
6. 時系列の前処理：対数変換

時系列の将来を予測する

時系列の過去の値を使うと簡単に予測できます



時刻 $n-1$ までの情報を使って求めた y_n の予測値を $y_{n|n-1}$ と表し，実際の観測値との差を予測誤差と呼びます．

どんな予測式を使うかが重要で，それによって予測の精度が変わってきます．

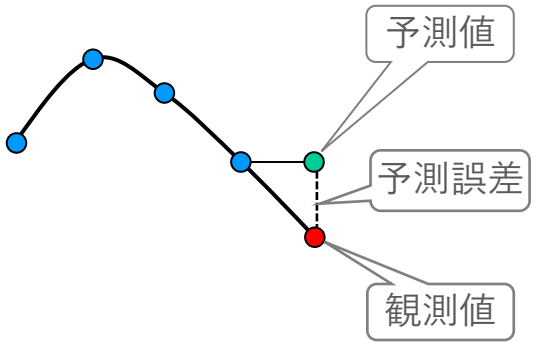
様々な予測式

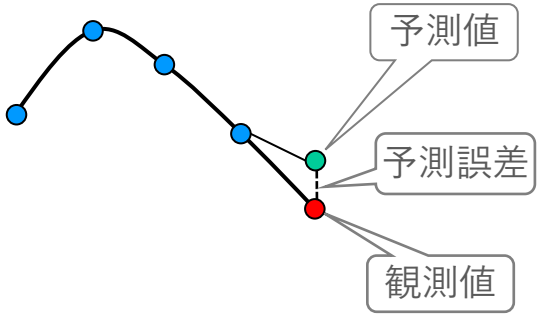
- ナイーブ予測　$y_{n|n-1} = y_{n-1}$
  前期と同じ値にするもの
  ランダムウォーク型

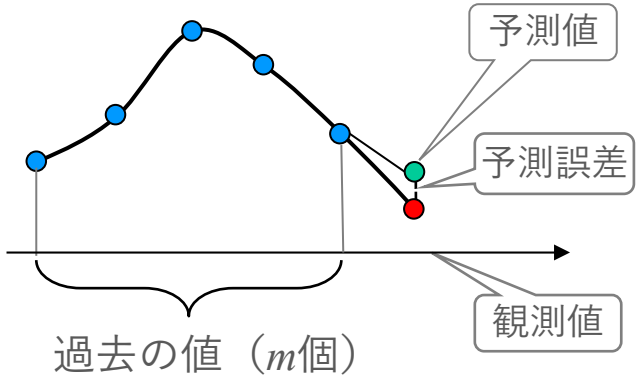
- 回帰型予測　$y_{n|n-1} = a_1 y_{n-1}$
  前期の値に適当な係数を掛けたもの
  マルコフ型

- 自己回帰型　$y_{n|n-1} = a_1 y_{n-1} + \cdots + a_m y_{n-m}$
  過去の値の重み付き平均
  いろいろな状況に対応した予測が
  できますが，次数 m や係数 $a_1, \ldots, a_m$
  をどうやって決めるかが問題です．

そのために以下ではARモデルを考えます．



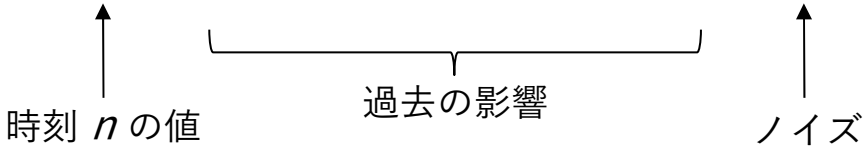




自己回帰モデル

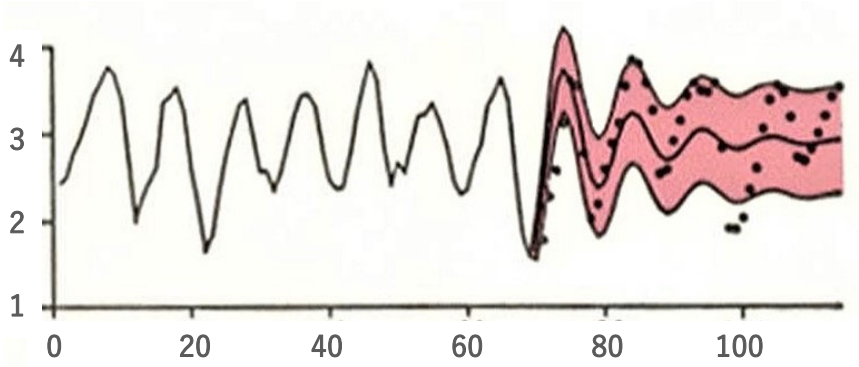
ARモデル（自己回帰モデル）は時系列の現在の値を過去の値で表現します.

$$y_n = a_1 y_{n-1} + \cdots + a_m y_{n-m} + v_n$$



- ノイズv_n は予測誤差と呼ばれ，平均 0, 分散 σ^2 の正規分布に従うと仮定します
- 時系列の過去の値だけを使って予測できます
- 予測値は $y_{n|n-1} = a_1 y_{n-1} + \cdots + a_m y_{n-m}$，予測誤差分散は σ^2
- 長期予測も簡単です（ただし，その誤差評価は少し面倒）

自己回帰モデルによる予測の例

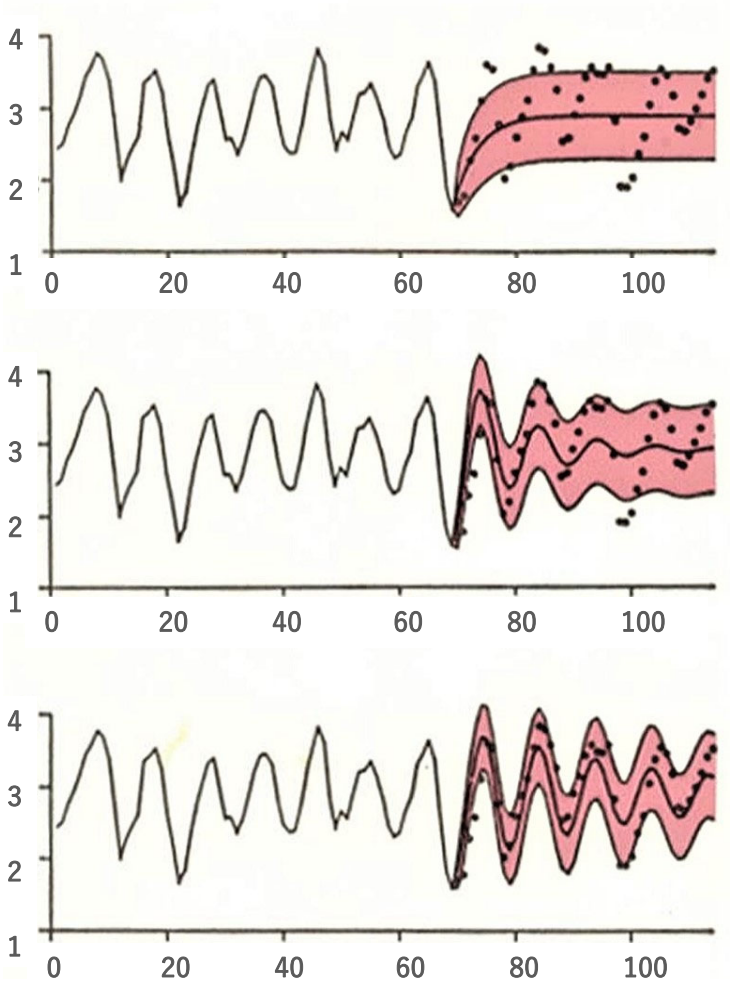


Canadian Lynx data

前半70個のデータを使って後半47個の時系列を予測した場合

- 最初の10点位は精度よく予測できています
- しかし，その後の予測はあまり良い結果が得られていません

モデルの次数によって予測の良さは変わる



$m=1$

$m=5$

$m=10$

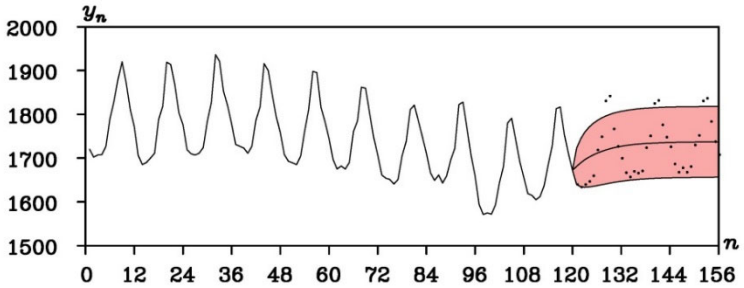
$$y_n = a_1 y_{n-1} + \cdots + a_m y_{n-m} + v_n$$

左図のようにモデルの次数 mを変えると予測結果が全く異なります.

- 高い次数ほど予測精度が良いとは限りません
- 適切な係数推定と次数選択が重要です
- 次数を選ぶ評価基準としてAIC, BICなどがあります

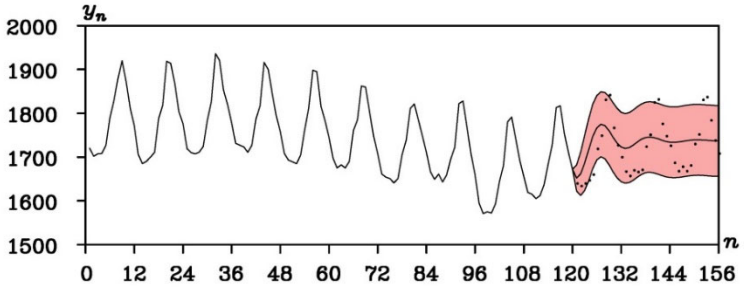
例：　長期予測　BIsallfoodデータ

$m=1$



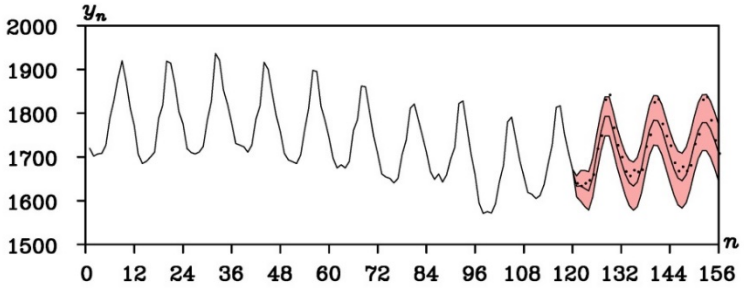
誤差幅に多くの実測値が入っていて間違いではないが，よい予測とはいえない．

$m=5$



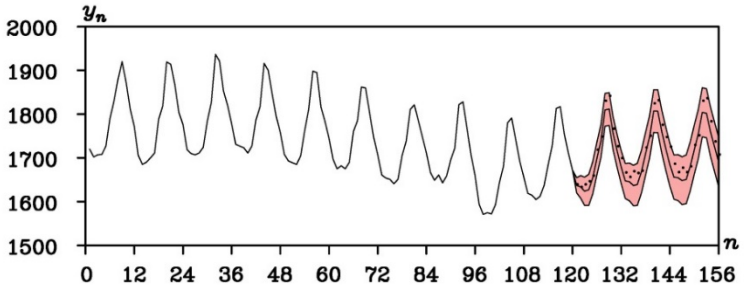
最初の半周期程度の予測は適当だが，それより先はm=1と大差ない．

$m=10$



36か月にわたってよい予測ができている．

$m=15$



さらに波形の特徴までよく再現できている．

ARモデルの同定（推定と次数選択）

- パラメータ推定

m 次のARモデルには自己回帰係数 $a_1, \ldots, a_m$ と予測誤差分散 σ^2 の $m+1$個の未知数（パラメータ）があります.

  - 自己回帰係数は予測誤差分散の期待値を最小化するYule-Walker法によって推定できます.
  - その他, 最小二乗法, Burg法, 最尤法, ベイズ法などがあります.

- 次数選択
  - 予測誤差の意味で最適な次数を選択するには情報量規準

$$\mathrm{AIC}_m = N(\log 2\pi\hat{\sigma}^2 + 1) + 2(m+1)$$

を最小にする m を探します．ただし，$\hat{\sigma}^2$ は予測誤差分散の推定値です.

時系列の前処理

- 複雑な時系列でも適切な前処理で分析が簡単になることがあります
- 以下のような前処理がよく利用されます
    - 変数変換
    - 階差（差分），季節階差
    - 前期比・前年比
    - 移動平均
    - 欠測値・異常値処理

＊ ここでは変数変換だけを取り上げます．階差，季節階差，移動平均は説明済みです．

＊ 4-7　データハンドリングの時系列版です．

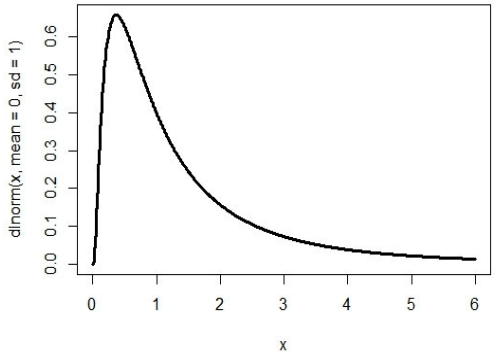
対数変換

複雑な時系列でも変数変換で分析が簡単になることがあります.

例：金額，個数，雨量など正値をとるデータの場合

- 分布が対象でない
- 平均がよい代表値でない
- 値が大きくなると変動も増える

などの問題があります。こんな時は対数変換 $y = \log x$ が有効です

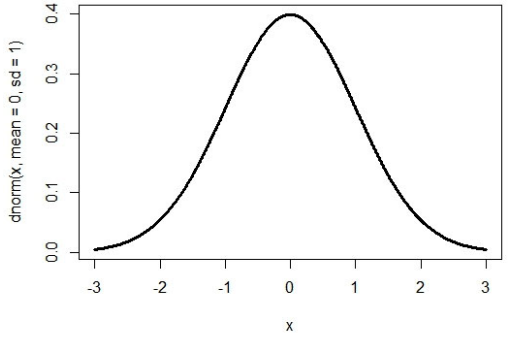


$y = \log x$

変換

逆変換

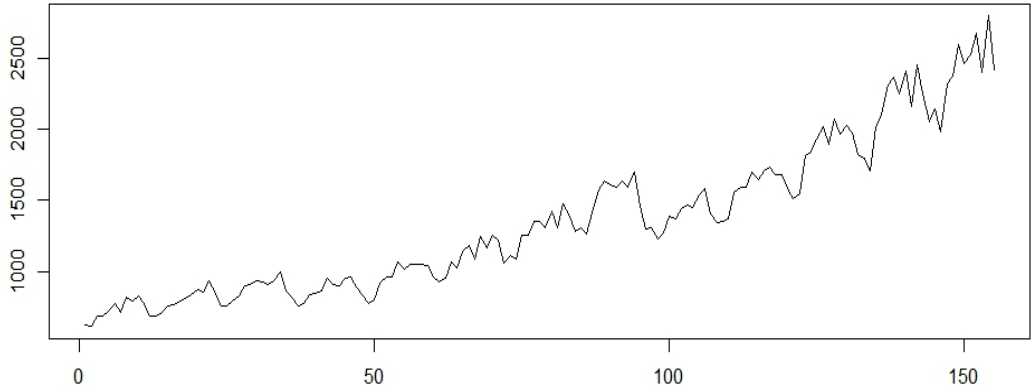
$x = e^{y}$



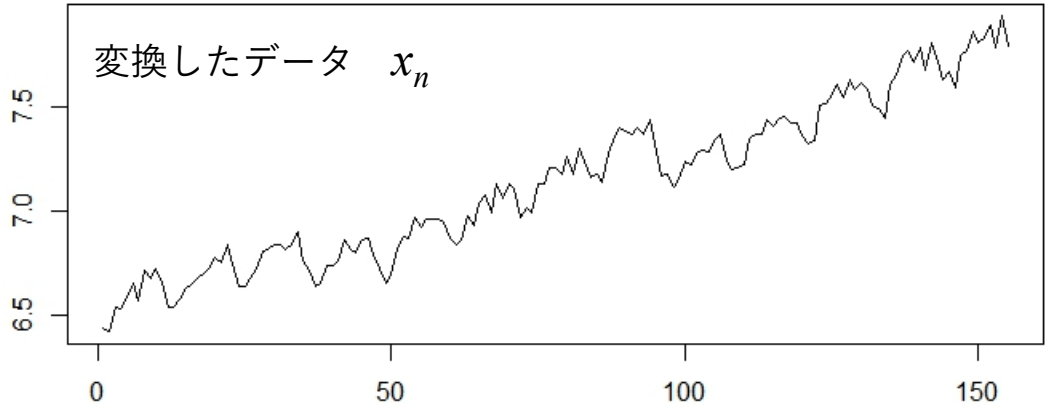
こんな分布が ⟶ 正規分布に！

もとの変数に戻せます ⟵ こちらの分析結果を

時系列の対数変換



$$x_n = \log y_n$$



経済時系列では，値が増加すると変動幅も大きくなるものが多い！

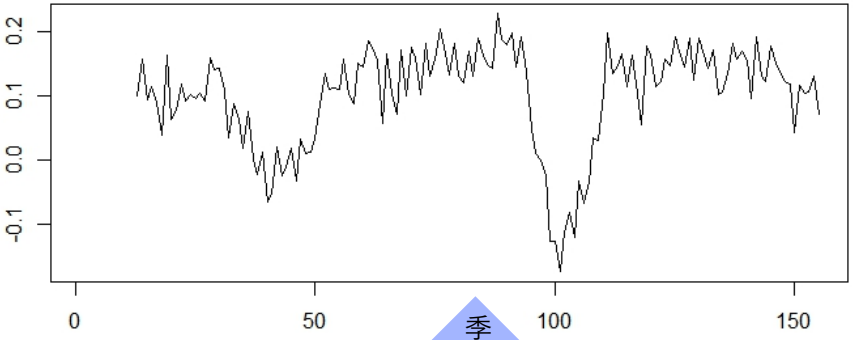
対数変換してみると・・・

- 変動幅がほぼ一定になる
- トレンドが直線的になる

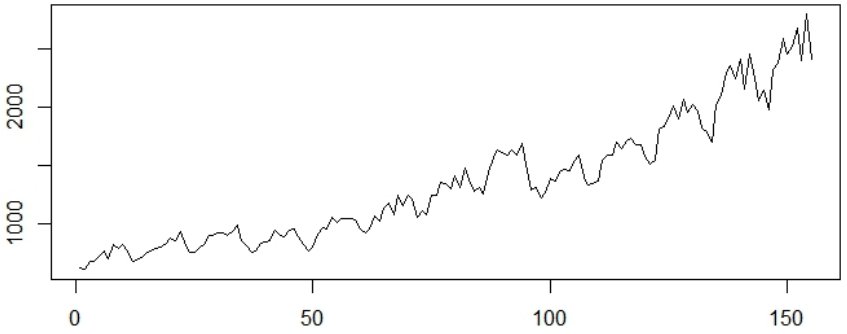
分析が楽になる

対数変換，階差，季節階差の影響まとめ

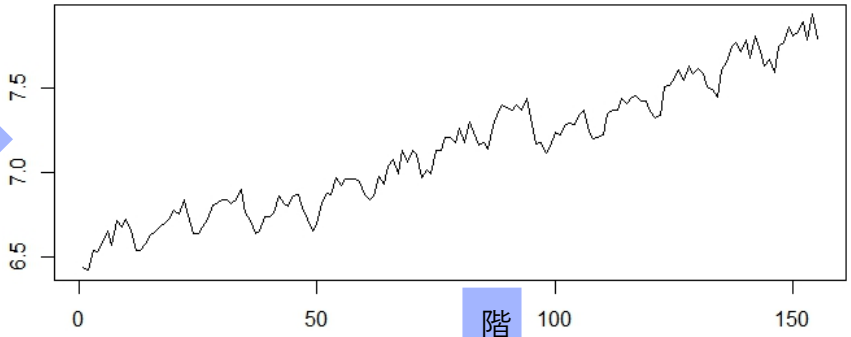
トレンドと季節変動が消えて
n=40と100付近の落ち込みが
顕著になりました



季節階差



対数



階差

トレンドは消えましたが
p=12の周期的変動は残っ
ています