L12 時系列データの分析—時系列解析

樋口さぶろお

龍谷大学 先端理工学部 数理・情報科学課程

理論物理学特論 L12(2021-12-21 Tue)

最終更新: Time-stamp: "2021-12-21 Tue 08:21 JST hig"

今日の目標

- 多変量データと時系列データの違いを説明できる
- 移動平均, 自己相関係数, コレログラムの意味を説明できる. Pandas で求められる.

L11-Q1

Quiz 解答:不偏推定量

X_i は独立同分布にしたがう.

1. $\mathrm{E}[\overline{X}] = \mathrm{E}[X_i]$ なので不偏推定量である. バイアスは 0.
2. $\mathrm{E}[\overline{X}'] = \mathrm{E}[X_i]$ なので不偏推定量である. バイアスは 0.
3. $\mathrm{E}[S^2] = \frac{n-1}{n}\mathrm{V}[X_i]$ なので不偏推定量でない. バイアスは $(\frac{n-1}{n} - 1)\sigma^2$.
4. (前略) 不偏推定量でない. (後略)

L11-Q2

Quiz 解答:ブートストラップ法

1. $\frac{1}{20}[8 + 8 + \cdots + 12] = 11$. バイアス $b = 11 - 10 = 1$.
2. $10 - (11 - 10) = 2 \times 10 - 11 = 9$.
3. $\frac{1}{20-1}[(8 - 11)^2 + \cdots + (12 - 11)^2] = \frac{24}{19}$.

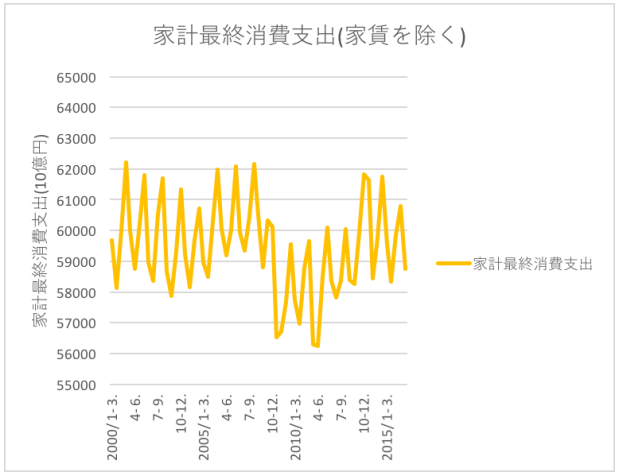
4. $0.05, 0.95$ quantile は $8, 12$. よって,
$10 - (12 - 10) < \theta < 10 + (8 - 10)]$.

ここまで来たよ

時系列解析 Time Series Analysis

時系列　時間 t に依存する量の列 $x(0), x(1), x(2), \ldots, x(t), \ldots$.
以前の値が, 今の値に影響. $x(t)$ $t = 0, 1, 2, 3, \ldots$ は独立でない.
例

- 特定の銘柄の毎日の株価のデータ
- 週ごとの売上のデータ
- 1 分おきの気温のデータ
- 1 年ごとの太陽黒点の個数のデータ
- 時刻 t のランダムウォーカーの座標 $X(t)$



時系列解析　時系列を解析する手法群 経済統計学でさかん.
目的 時系列を再現する. $t \leq T$ のデータから $t > T$ を予測する.

標本 (データ) を解析 $\rightarrow$ **時系列モデルを作成** $\rightarrow$ 再現・予測

ランダムウォーク　　確率モデル (3Q4)

時系列データの形式

多変量時系列データ $\boldsymbol{x}(t)$.

日時	気温	気圧
2021-12-18 Sat 10:00	10.1	1010
2021-12-18 Sat 11:00	10.5	1005
⋮	⋮	⋮

日時の形式はややこしい, けどしばらく放っておく
日時には実質的な意味がある. 日時を消して順序を変えてはいけない.
日時 i のデータは日時 $j(<i)$ のデータの影響を受けることがある.
しばらく, 日時の間隔はすべて同じ, 日時順に並んでる, として, 日時を表示しないことがある
多変量データ

選手番号	100m 走	棒高跳び
1	10.1	2.30
2	10.4	2.12
⋮	⋮	⋮

選手番号は便宜上のもので, 選手番号を振り直したデータも同じ意味.
(同じ分布からとってきたデータ点かもしれないが) 選手 i のデータが選手 j のデータの影響を受けるようなことはない.

時系列の 3 要素

現実の時系列は次の 3 つの重ね合わせになっていることが多い

$$x(t) = \boxed{\text{トレンド}} + \boxed{\text{周期的変動}} + \boxed{\text{ランダム成分}}$$

- トレンド (長期的傾向) 期間を通して時間に比例して増減する傾向.
  —過的な増減
- 短い周期の周期的変動 季節, 週, 月, 年

フーリエ級数解析, フィルタ (パターン情報処理)

- ランダム成分 (ノイズ) 時刻ごとに独立な確率変数

定常な時系列

(標本ナントカ) の意味で, 日時のどの一部分を取り出しても「本質的に」同じであるとき, 定常 (stationary) という
トレンドがあれば非定常 (氷河期と今は同じでない)
周期的変動があるなら非定常 (夏と秋は同じでない, 昼と夜は同じでない)
トレンドや周期的変動を引いておくと定常になる. トレンドや周期的変動を移動平均で消すと定常になる.
近いうちに, 母ナントカの立場で定常を明確に定義します.

ここまで来たよ

移動平均 Moving Average

時系列 $x(t)$ から平滑化 (smoothing) した別の時系列 $y(t)$ を作る手法

$(2\ell+1)$ 次の移動平均

$$y_{2\ell+1}(t) = \frac{1}{2\ell+1} \sum_{t'=t-\ell}^{t+\ell} x(t')$$

.

$x(0)$ $x(1)$ $x(2)$ $x(3)$ $x(4)$ $x(5)$ $x(6)$ $x(7)$ $x(8)$ $x(9)$ $x(10)$ x

2ℓ 次の移動平均

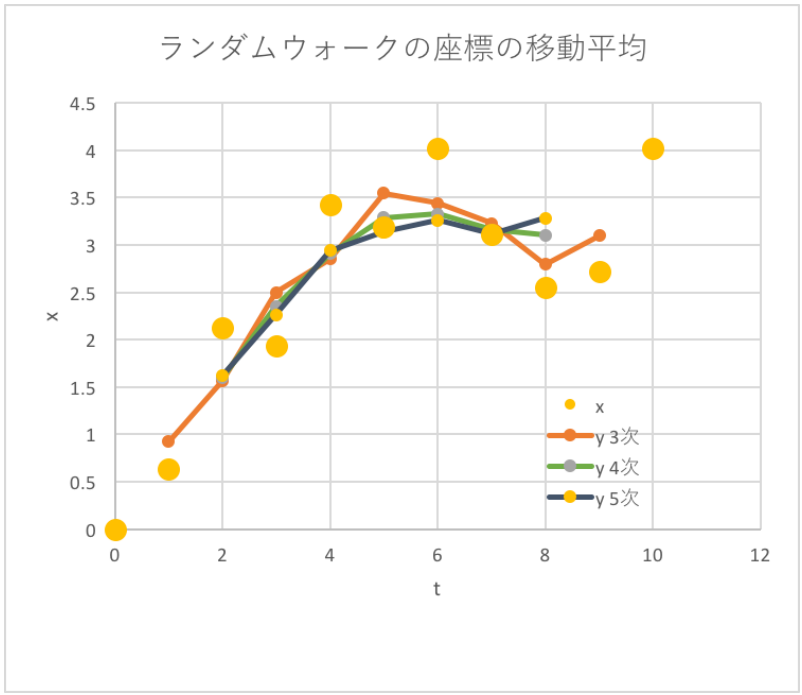
$$y_{2\ell}(t) = \frac{1}{2\ell}\left(\tfrac{1}{2}\cdot x(t-\ell+1) + x(t-\ell+2) + \cdots + x(t) + \cdots + x(t+\ell-2) + \tfrac{1}{2}x(t+\ell-1)\right)$$

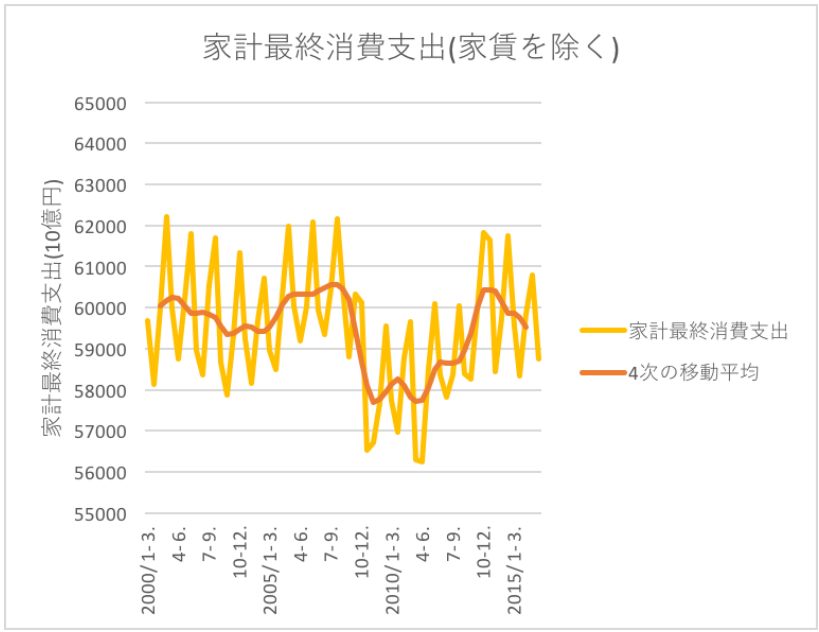
$x(0)$ $x(1)$ $x(2)$ $x(3)$ $x(4)$ $x(5)$ $x(6)$ $x(7)$ $x(8)$ $x(9)$ $x(10)$ x

離散フーリエ級数の第 0 項 th-d12-tsa.ipynb

移動平均の性質

- 次数が高くなるほど滑らかになる
- 元のデータの「真ん中へん」を通る.
- 時間帯の端までは描けない





移動平均の性質 2

- トレンド (長期的傾向) 期間を通して時間に比例して増減する傾向.
  ー過的な増減 $\to$ 移動平均ではっきり見えるようになる
- 短い周期の周期的変動 季節, 週, 月, 年 $\to$ (周期くらいの) 移動平均で消える. (もとのデータ)−(移動平均) ではっきり見える フーリエ級数解析, フィルタ (パターン情報処理)
- ランダム成分 (ノイズ) 時刻ごとに独立な乱数 $\to$ 移動平均で小さくなる

ラグ ℓ の後方移動平均 $y(t) = \frac{1}{\ell}\sum_{t'=t-\ell+1}^{t} x(t)$
ラグ ℓ の前方移動平均 $y(t) = \frac{1}{\ell}\sum_{t'=t}^{t+\ell-1} x(t)$

L12-Q1

Quiz(移動平均)

次の時系列データから, 3,4 次の移動平均を求めよう.

t	0	1	2	3	4	5	6	7	8	9
x	1.8	-1.6	2.6	-1.2	3.0	-1.2	3.4	-0.8	3.8	0.0
y_3										
y_4										

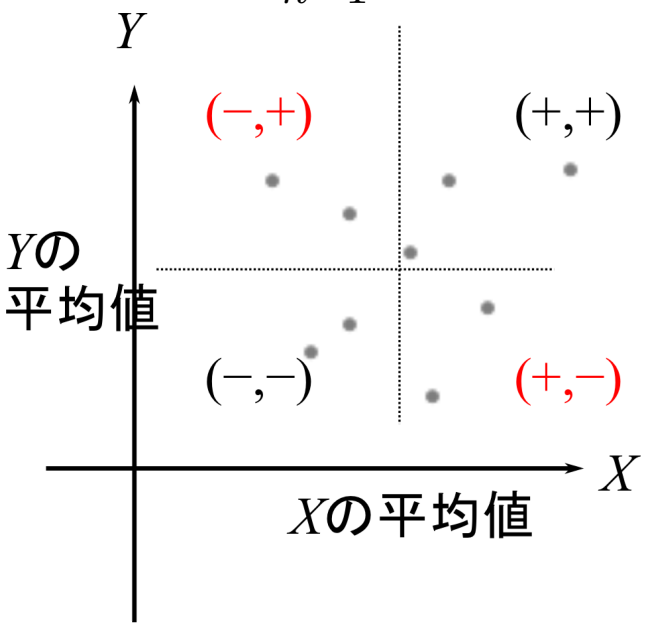
ここまで来たよ

復習:標本共分散と標本相関係数

標本共分散 (covariance)

$$x, y \text{ の共分散 } C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \times (y_i - \overline{y})$$

(不偏なら $\frac{1}{n-1}$)

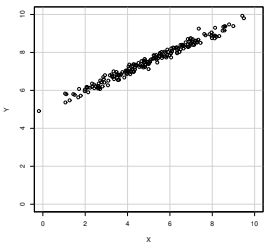


$$(+,-) = (x_i - \overline{x}\text{の符号}, y_i - \overline{y}\text{の符号}).$$

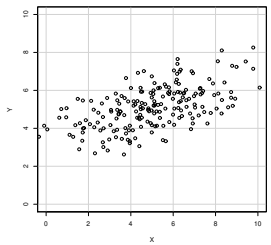
相関係数

$$\text{標本相関係数 } r = \frac{C_{xy}}{s_x s_y}$$

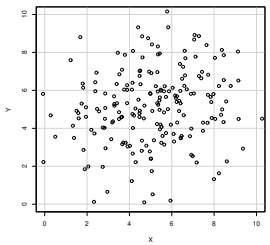
s_x, s_y :標本標準偏差
分子分母で $\frac{1}{n}$, 不偏 $\frac{1}{n-1}$ をそろえる.



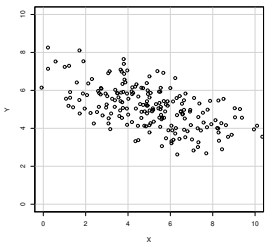
強い正の相関
$r = 0.99$



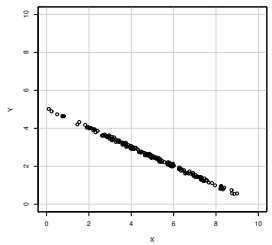
弱い正の相関
$r = 0.55$



無相関
$r = 0$



弱い負の相関
$r = -0.55$



強い負の相関
$r = -0.99$

相関
‘正の相関’: x が大きい $\Leftrightarrow$ y が大きい
‘負の相関’: x が大きい $\Leftrightarrow$ y が小さい
強い/弱い: 傾向がはっきりしている/していない

2 変量データ

標本自己共分散，標本自己相関係数

k 次の標本自己共分散，標本自己相関係数

時間 t を，ラグ k だけずらした $x(t), x(t-k)$ を 2 変量データだと思って，標本共分散，標本相関係数を考えたもの

$k=1$ の例

x	y
$x(1)$	—
$x(2)$	$x(2-1)$
$x(3)$	$x(3-1)$
$\vdots$	$\vdots$
$x(T-1)$	$x(T-1-1)$
$x(T)$	$x(T-1)$
—	$x(T)$

k の例

x	y
$x(1)$	—
$\vdots$	
$x(k+1)$	$x(1)$
$\vdots$	$\vdots$
$x(T)$	$x(T-k)$
$\vdots$	
—	$x(T)$

kj 次の標本自己共分散 autocovariance

$$C(k) = \frac{1}{T-k} \sum_{t=k+1}^{T} (x(t) - \overline{x})(x(t-k) - \overline{x})$$

ただし標本平均値$\overline{x} = \dfrac{1}{T-k} \displaystyle\sum_{t=k+1}^{T} x(t)$

kj 次の標本自己相関係数 autocorrelation

$$r(k) = \frac{\text{自己共分散}}{\sqrt{\text{分散}}\sqrt{\text{分散}}} = \frac{C(k)}{C(0)}$$

$C(0)$ は $x(t)$ をサイズ T の標本と思ったときの分散.

コレログラム correlogram

横軸 ラグ k, 縦軸 k 次の自己相関係数 のグラフ.

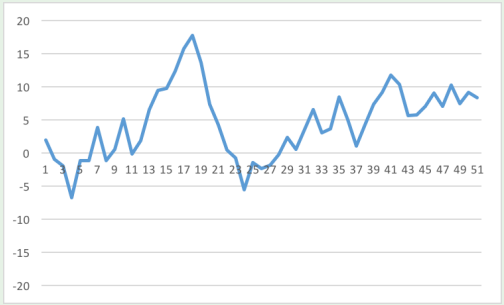
- 多くの定常モデル (自己回帰モデルなど) では, k が大きいほど (遠い時刻ほど), 標本自己相関係数 $r(k)$ の絶対値は小さくなる.
- ランダムウォークのとき, $r(k) =$ 一定.

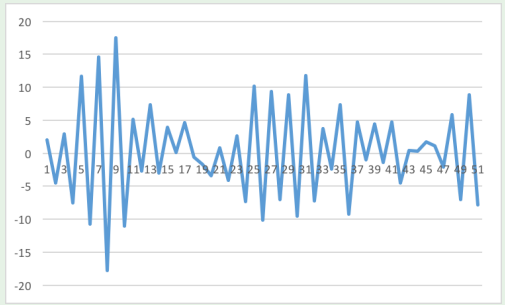
th-d12-tsa.ipynb

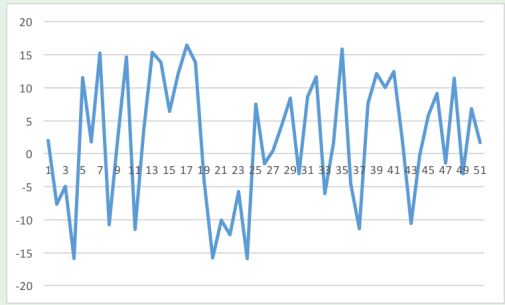
L12-Q2

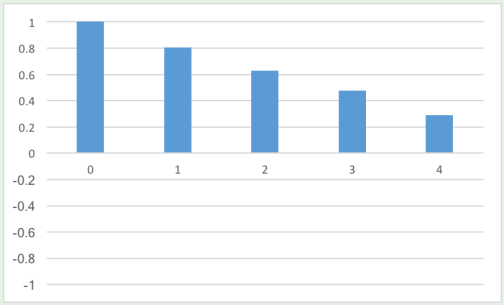
Quiz(コレログラム)

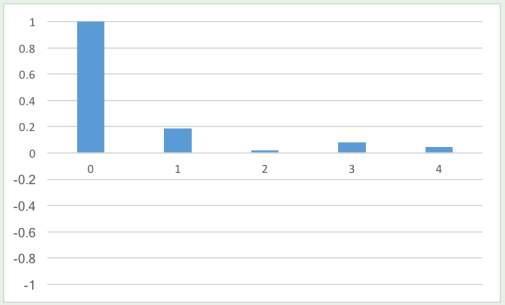
次の時系列データと，コレログラムの間の対応をつけよう．

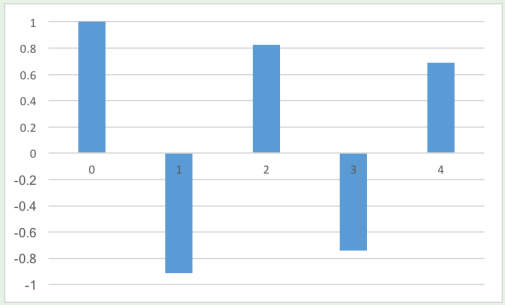
1.


2.


3.


a.


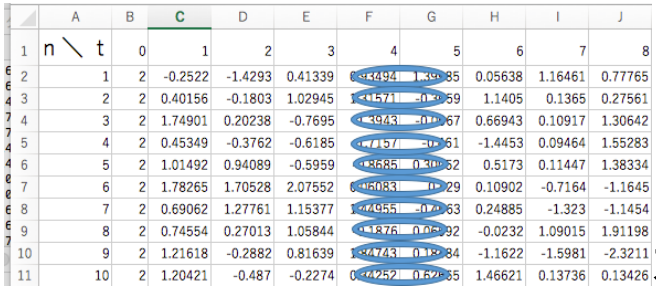
b.


c.


定常な時系列データに対する標本ナントカ

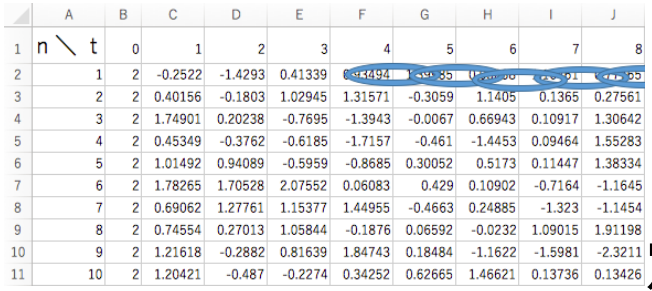
‘時系列分析’ では, 定常過程について, 1 個の時系列データを, 一定の長さに分割して複数個のデータからなる標本のように扱うことが行われる.

横:t 縦:標本内のデータ番号. 標本自己相関係数を求めるとき



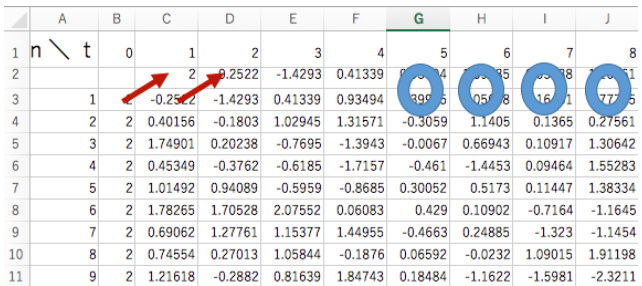
n＼t	0	1	2	3	4	5	6	7	8
1	2	-0.2522	-1.4293	0.41339	[illegible]		0.05638	1.16461	0.77765
2	2	0.40156	-0.1803	1.02945	[illegible]		1.1405	0.1365	0.27561
3	2	1.74901	0.20238	-0.7695	[illegible]		0.66943	0.10917	1.30642
4	2	0.45349	-0.3762	-0.6185	[illegible]		-1.4453	0.09464	1.55283
5	2	1.01492	0.94089	-0.5959	[illegible]		0.5173	0.11447	1.38334
6	2	1.78265	1.70528	2.07552	[illegible]		0.10902	-0.7164	-1.1645
7	2	0.69062	1.27761	1.15377	[illegible]		0.24885	-1.323	-1.1454
8	2	0.74554	0.27013	1.05844	[illegible]		-0.0232	1.09015	1.91198
9	2	1.21618	-0.2882	0.81639	[illegible]		-1.1622	-1.5981	-2.3211
10	2	1.20421	-0.487	-0.2274	[illegible]		1.46621	0.13736	0.13426

本当はこういう標本が欲しい



n＼t	0	1	2	3	4	5	6	7	8
1	2	-0.2522	-1.4293	0.41339	[illegible]				
2	2	0.40156	-0.1803	1.02945	1.31571	-0.3059	1.1405	0.1365	0.27561
3	2	1.74901	0.20238	-0.7695	-1.3943	-0.0067	0.66943	0.10917	1.30642
4	2	0.45349	-0.3762	-0.6185	-1.7157	-0.461	-1.4453	0.09464	1.55283
5	2	1.01492	0.94089	-0.5959	-0.8685	0.30052	0.5173	0.11447	1.38334
6	2	1.78265	1.70528	2.07552	0.06083	0.429	0.10902	-0.7164	-1.1645
7	2	0.69062	1.27761	1.15377	1.44955	-0.4663	0.24885	-1.323	-1.1454
8	2	0.74554	0.27013	1.05844	-0.1876	0.06592	-0.0232	1.09015	1.91198
9	2	1.21618	-0.2882	0.81639	1.84743	0.18484	-1.1622	-1.5981	-2.3211
10	2	1.20421	-0.487	-0.2274	0.34252	0.62665	1.46621	0.13736	0.13426

定常ならこれでもいいじゃん



n＼t	0	1	2	3	4	5	6	7	8
1	2	-0.2522	-1.4293	0.41339	0.93494				
2	2	0.40156	-0.1803	1.02945	1.31571	-0.3059	1.1405	0.1365	0.27561
3	2	1.74901	0.20238	-0.7695	-1.3943	-0.0067	0.66943	0.10917	1.30642
4	2	0.45349	-0.3762	-0.6185	-1.7157	-0.461	-1.4453	0.09464	1.55283
5	2	1.01492	0.94089	-0.5959	-0.8685	0.30052	0.5173	0.11447	1.38334
6	2	1.78265	1.70528	2.07552	0.06083	0.429	0.10902	-0.7164	-1.1645
7	2	0.69062	1.27761	1.15377	1.44955	-0.4663	0.24885	-1.323	-1.1454
8	2	0.74554	0.27013	1.05844	-0.1876	0.06592	-0.0232	1.09015	1.91198
9	2	1.21618	-0.2882	0.81639	1.84743	0.18484	-1.1622	-1.5981	-2.3211

Excel 的にはこうやると楽

ここまで来たよ

12 ブートストラップ法

12 時系列データの分析—時系列解析
- 時系列データ
- 移動平均
- 自己相関係数
- 時系列データの変換

時系列データ=数列

$a_n \leftrightarrow x(t)$.

階差数列

$y(t) = x(t) - x(t-1)$.

増加率

$y(t) = x(t)/x(t-1)$, $y(t) = x(t)/x(t-1) - 1$.

等比数列 $rightarrow$ 等差数列

$x(t) = ar^t \to y(t) = \log x(t) = \log a + t \log r$.

増加率 $\to$ 階差

$t \log r$: トレンド