日本行動計量学会セミナー

切片と傾きが時間に応じて変化する回帰モデル

線形回帰モデルの注意点

時系列データに対して回帰分析を行う際
どのようなことに注意が必要だろうか

時系列データ

時系列データ

時点ごとに得られた測定値を、時間の順に並べたデータ
「データの並び順」に意味があるのが特徴

- 東京都の毎日の気温の推移
- 1秒ごとにとられた株価
- 毎月の売り上げデータ
- 30分ごとの電力需要

時系列データは、状態空間モデルを使うことで、
柔軟にモデル化できる

線形回帰モデル

<u>回帰分析によって得られた数式の例</u>

ビールの売り上げ = 22.79 + 気温 × 0.69 + 誤差

Y_i　　　　　　　　x_i

平均が0である
正規分布に従うと仮定

誤差は、互いに独立な正規分布を想定

独立

独立性のイメージ

1つ目の確率変数をX_1と、2つ目の確率変数をX_2とする
条件X_1があってもなくても、
X_2の確率分布が変わらないならX_1, X_2は独立

例えば……
1つ目のデータの残差が大きかった（X_1が大きい）か、
残差が小さかった（X_1が小さい）かは、
2つ目のデータの残差の大きさX_2に影響を与えない

自己相関

自己相関があるなら、
独立とは言えない

自己相関とは

時点同士の関連性を探る指標
「1時点前」と「今の時点」での相関をとるイメージ
(厳密な計算は微妙に異なるが省略)
正の自己相関と負の自己相関がある

正

- 昨日の売り上げが多ければ今日も多い
- 昨日の売り上げが少なければ今日も少ない

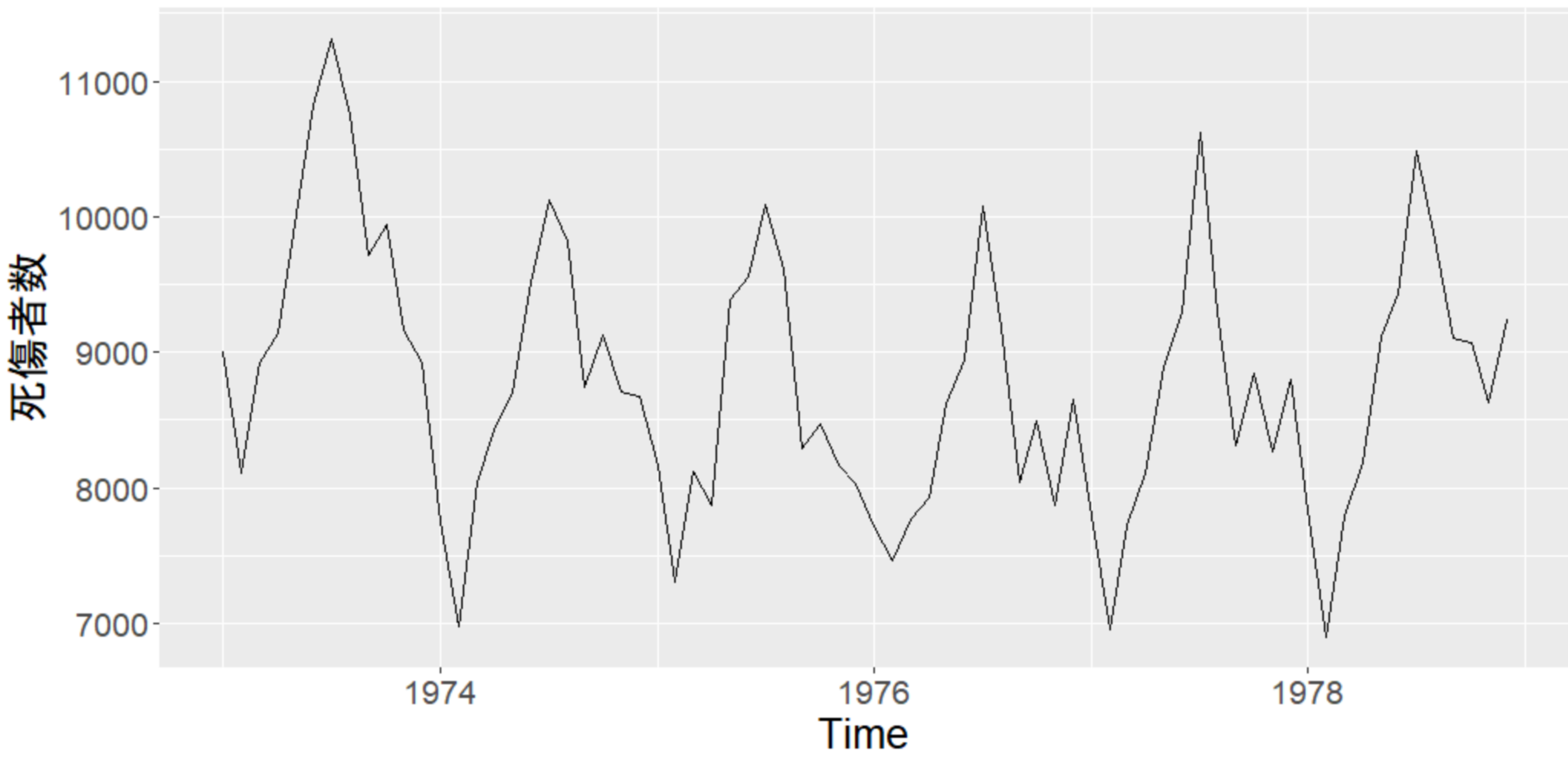
負

- 昨日の売り上げが多ければ今日は少ない
- 昨日の売り上げが少なければ今日は多い

自己相関

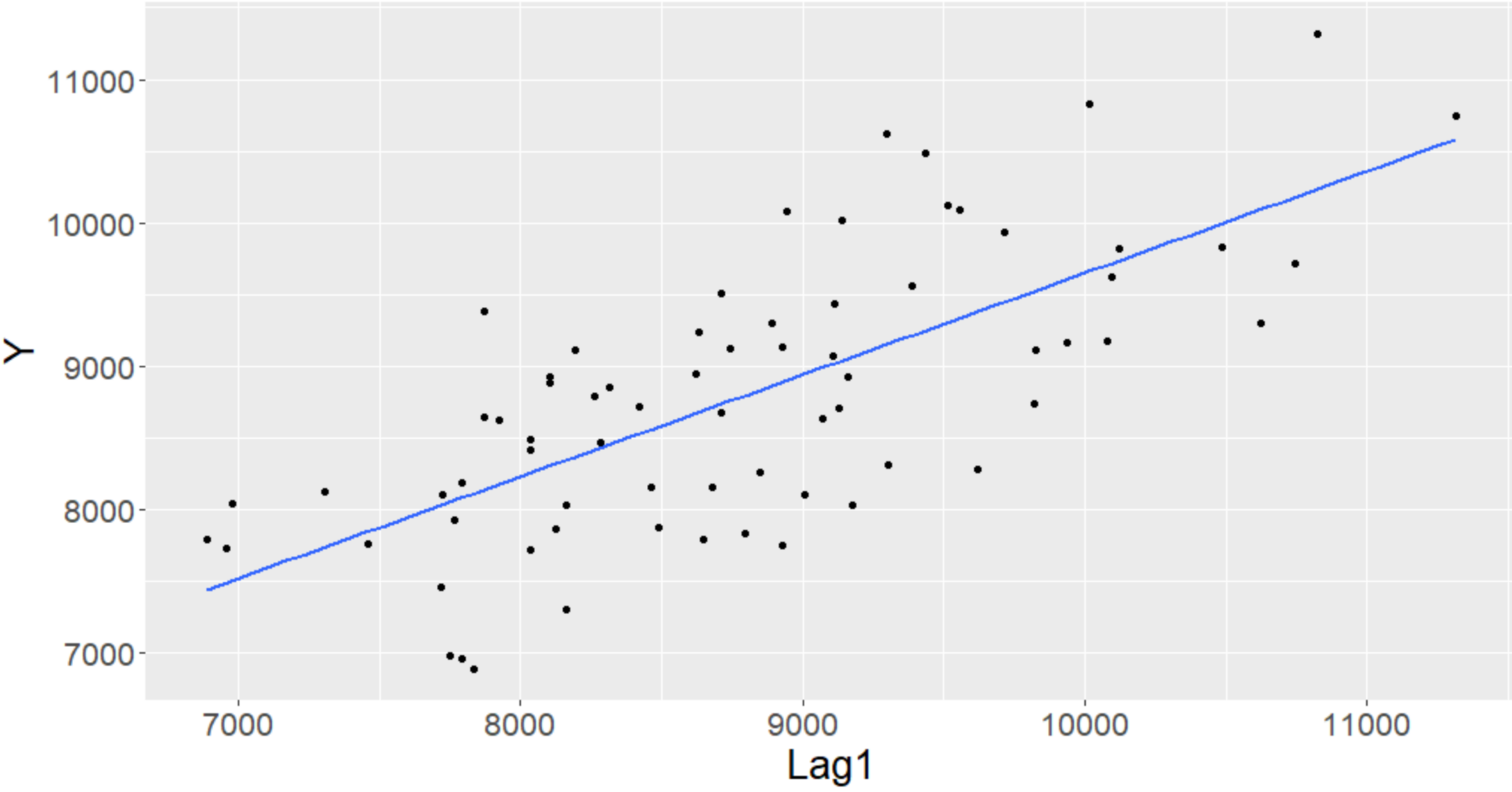
P. J. Brockwell and R. A. Davis (1991)
Time Series: Theory and Methods. Springer, New York.

1973年1月から1978年12月までの
アメリカの交通事故死傷者数の時系列データ



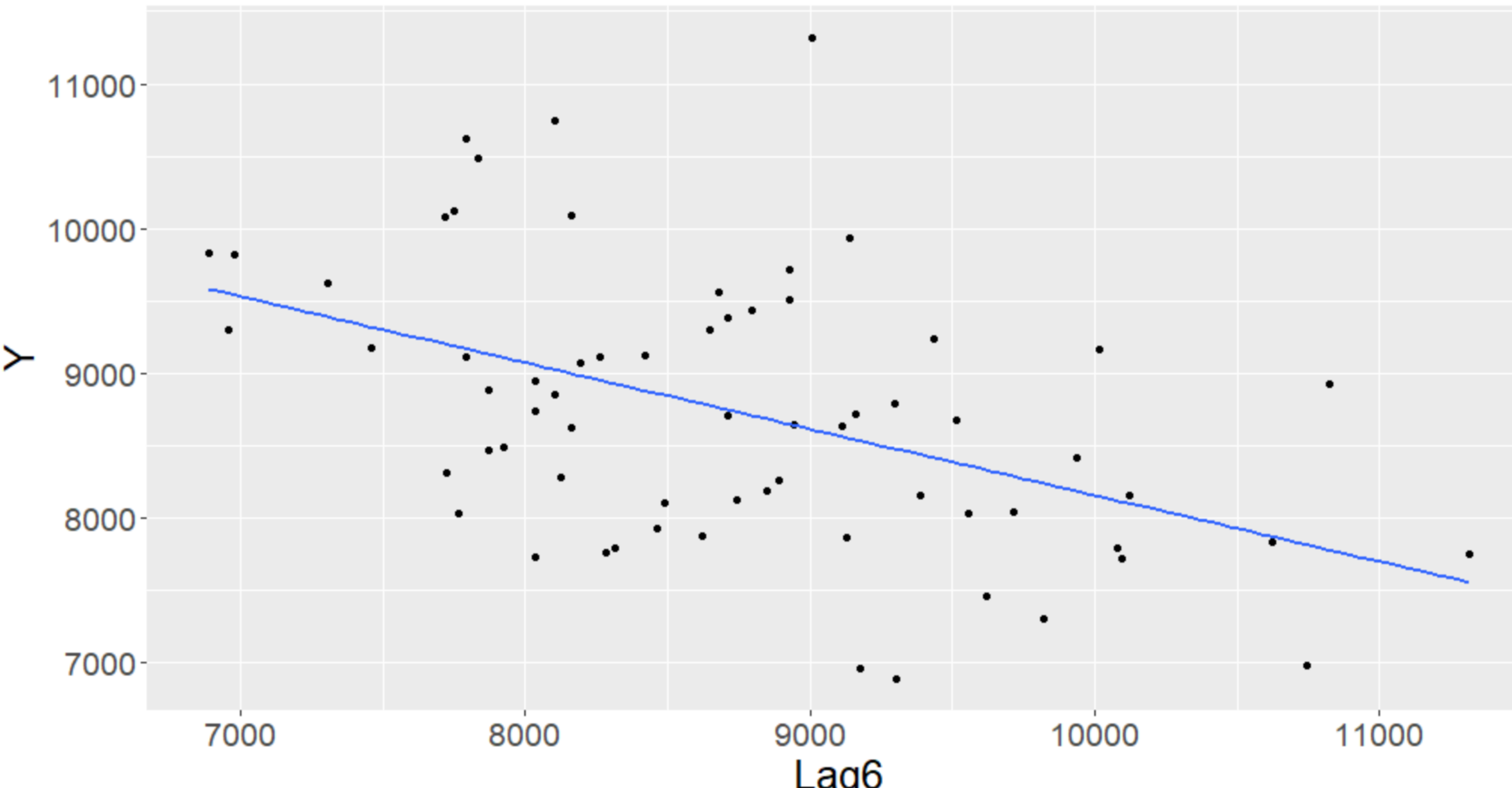
自己相関

横軸が「先月」　縦軸が「今月」の死傷者数のグラフ
→先月に死傷者数が多ければ今月も多い正の自己相関



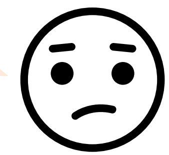
自己相関

横軸が「6か月前」　縦軸が「今月」の死傷者数のグラフ
→半年前に死傷者数が多ければ今月は少ない負の自己相関



回帰分析の注意点

残差に正の自己相関があると
回帰分析の結果は、どうなるだろうか



残差の自己相関のイメージ

大きな残差が出たら、次も大きな残差が出やすい
小さな残差が出たら、次も小さな残差が出やすい

回帰分析の注意点

自己相関のイメージ：パチンコの例

どちらのほうがデータ（お客の儲け）のばらつきが大きいか

儲け同士に相関は無い

当たり始めたら、ずっと当たる
外れ始めたら、ずっと外れる

回帰分析の注意点

自己相関のイメージ：パチンコの例

どちらのほうがデータ（お客の儲け）のばらつきが大きいか

儲け同士に相関は無い

儲かった人とそうでない人の差が激しい
→ばらつきが大きい！

当たり始めたら、ずっと当たる
外れ始めたら、ずっと外れる

回帰分析の注意点

残差が互いに独立ではない場合

残差同士に正の自己相関がある場合、
想定しているよりも残差のばらつきが大きくなる！
→統計的仮説検定の結果などにも影響が出るので注意

ランダムウォーク過程の初歩

ホワイトノイズ

ホワイトノイズ

予測できない純粋な「ノイズ」のこと

ホワイトノイズの特徴

平均　　　　　：ゼロ
分散　　　　　：一定
自己相関　　　：ゼロ

平均0分散σ^2の正規分布に従うノイズを
仮定することが多い（正規ホワイトノイズ）

ホワイトノイズとランダムウォーク

iid（独立で同一な確率分布）

ホワイトノイズ（自己相関無し）よりも、
より厳しい条件である。

正規ホワイトノイズ

正規分布に従うホワイトノイズ
正規ホワイトノイズはiidである

ランダムウォーク

iid系列（正規ホワイトノイズなど）の累積和。
独特の動きをするので、さまざまな場所で登場

**ランダムウォーク系列と
正規ホワイトノイズ系列の違いは？**

ランダムウォーク過程の初歩

定常過程

性質が一定で、時間的に変化しないもの

非定常過程

性質が時間と共に変化するもの

ランダムウォーク過程の初歩

<u>定常過程の性質1</u>

平均値や分散が一定である

→ トレンドなどが存在しない

<u>定常過程の性質2</u>

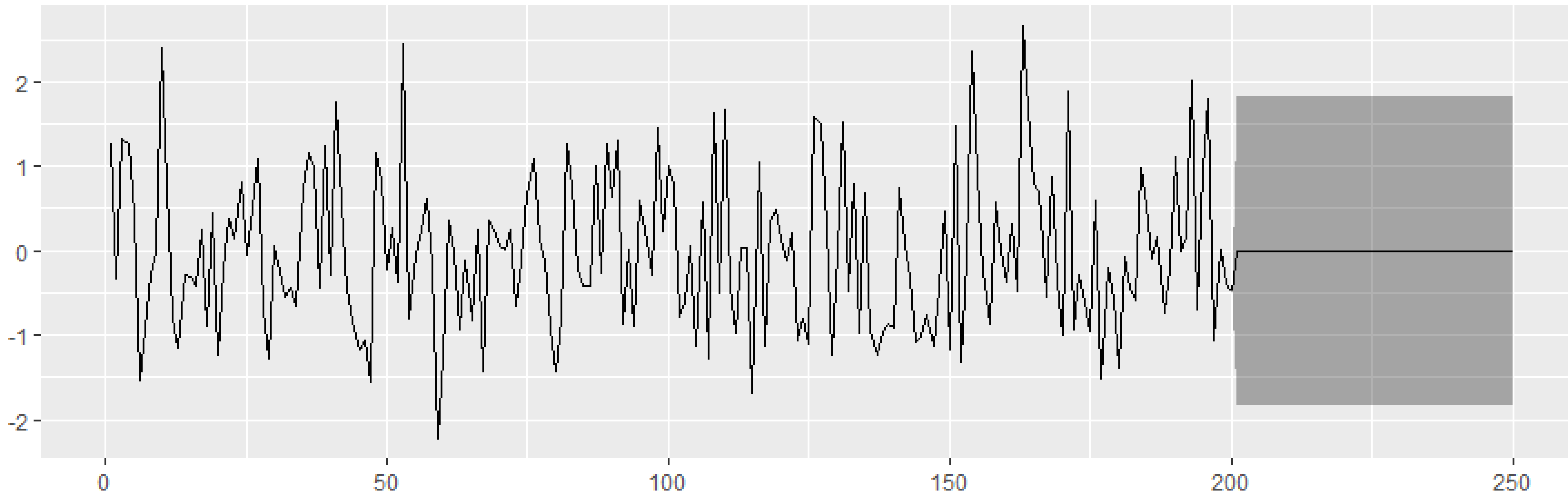
自己相関は時間差のみに依存

→ 2000年代における1日差の相関関係

＝

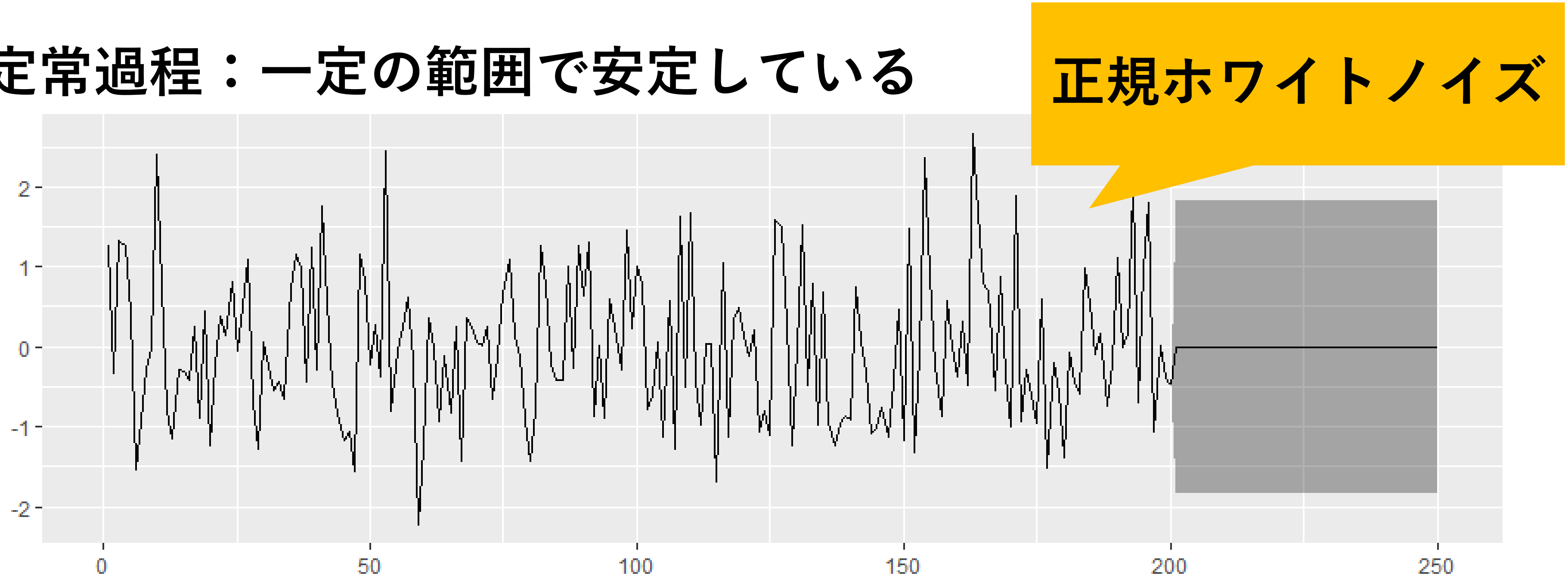
1950年代における1日差の相関関係

定常過程：一定の範囲で安定している

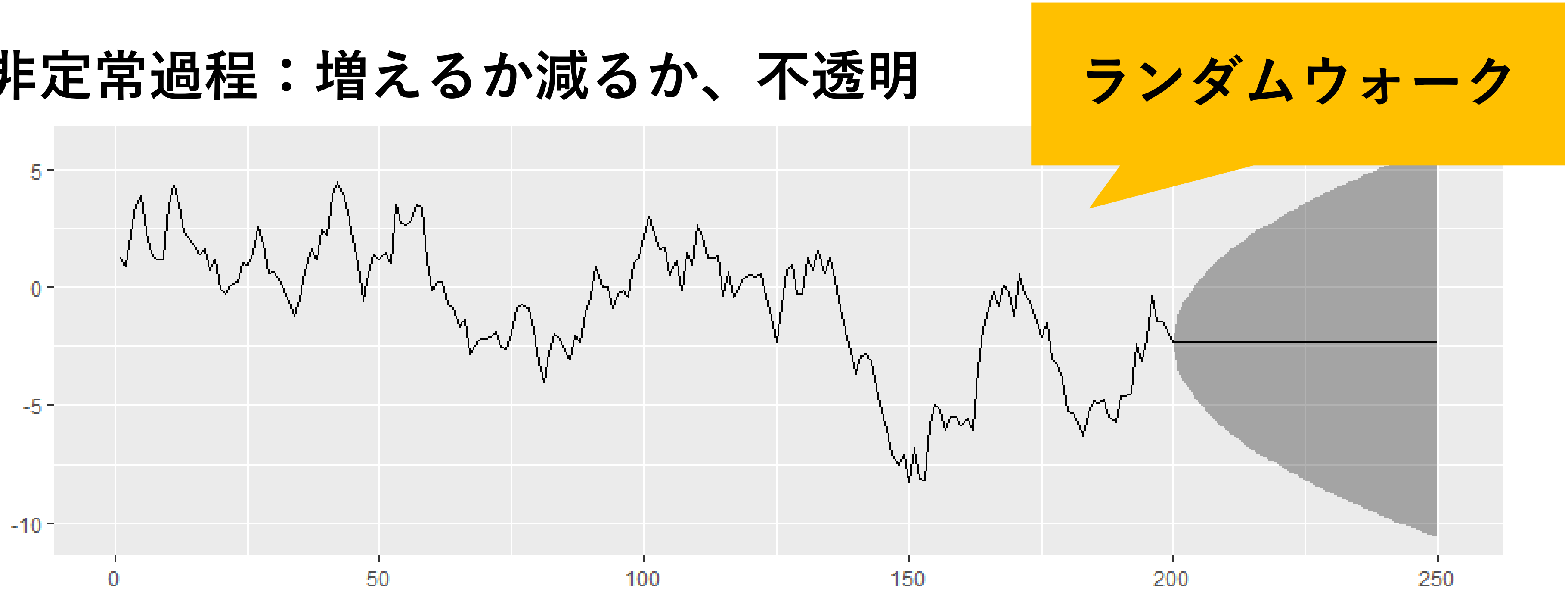


非定常過程：増えるか減るか、不透明

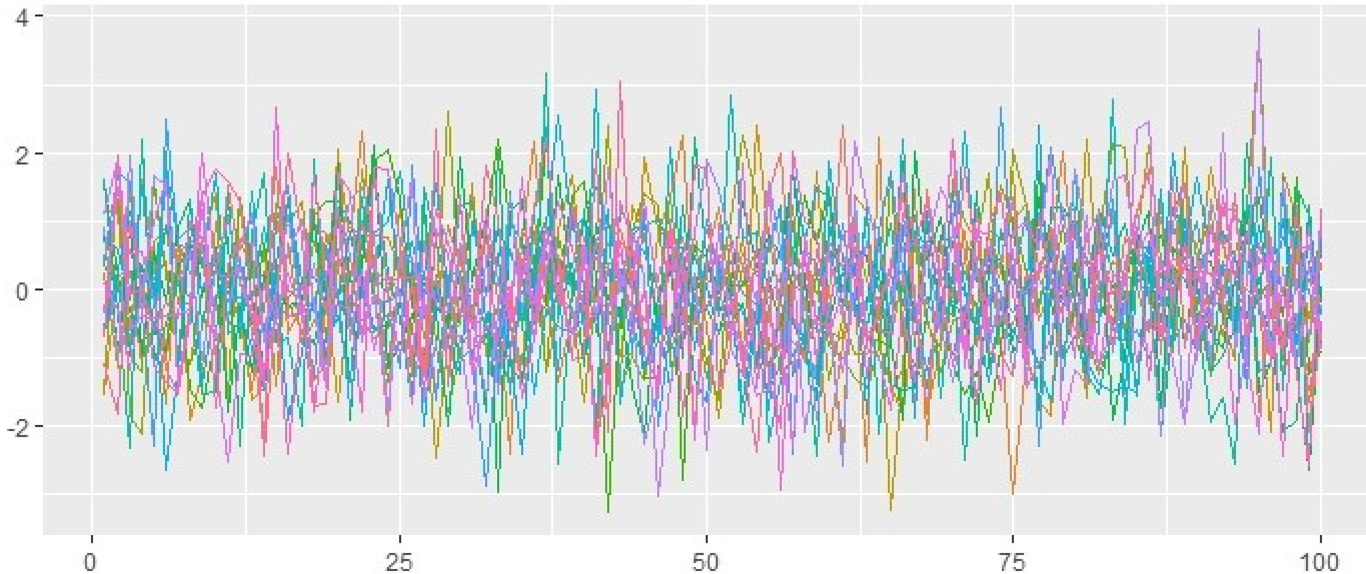
定常過程：一定の範囲で安定している



非定常過程：増えるか減るか、不透明

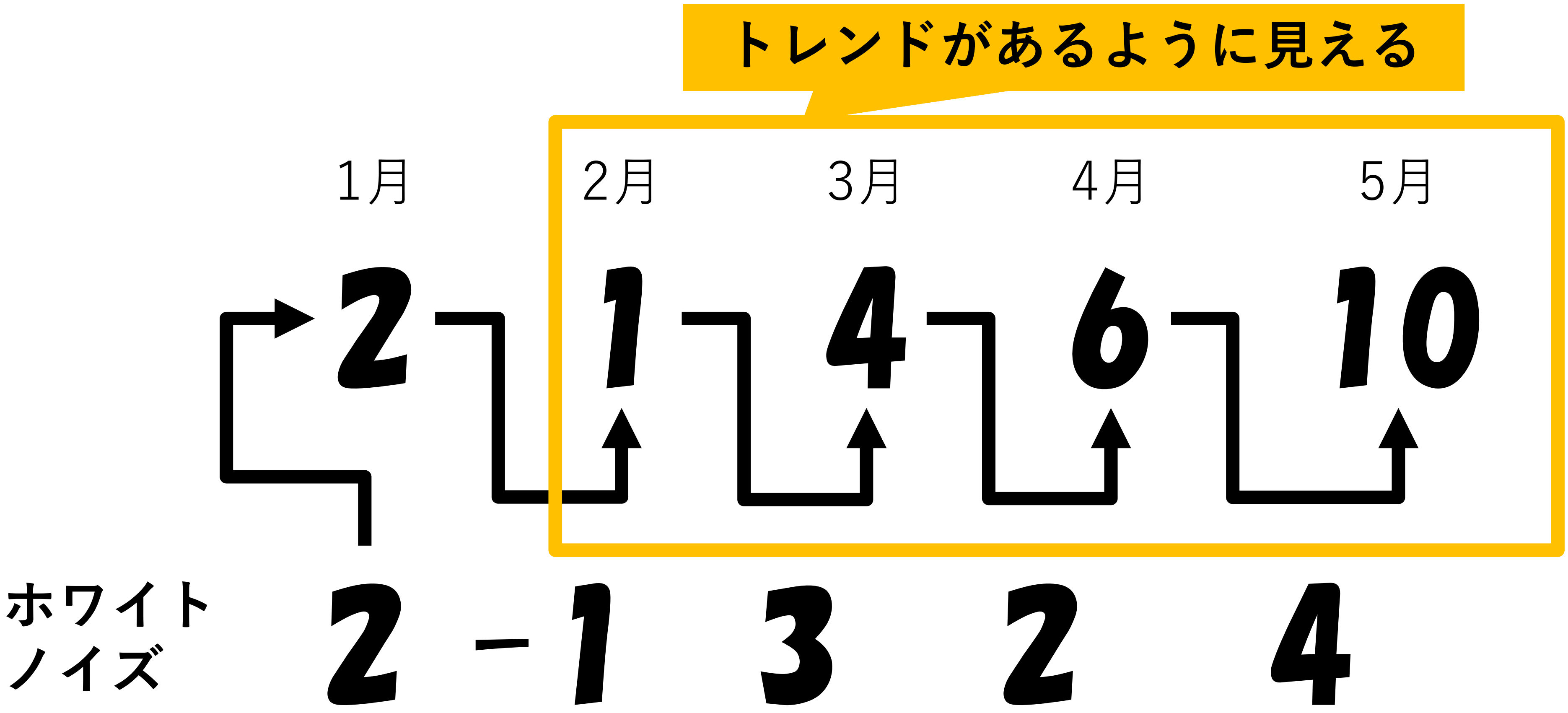


定常過程（正規ホワイトノイズ）：20系列をシミュレート



非定常過程（ランダムウォーク）：20系列をシミュレート

非定常過程の例：ランダムウォーク



正規ホワイトノイズの累積和は非定常

ランダムウォーク過程の初歩

ランダムウォーク(RW)

誤差が累積するので、分散がどんどん増える
→非定常過程

正規ホワイトノイズ(WN)

分散は一定
→定常系列

ランダムウォーク過程の初歩

ランダムウォーク系列と自己相関

iid系列（正規ホワイトノイズなど）の累積和
ランダムウォークは「1時点前の自分の値」を含む

ランダムウォーク(RW)のシミュレーション

1日目のRW　＝ WN①

2日目のRW　＝ WN① ＋ WN②

3日目のRW　＝ WN① ＋ WN② ＋ WN③

4日目のRW　＝ WN① ＋ WN② ＋ WN③ ＋ WN④

ランダムウォーク過程の初歩

ランダムウォーク系列と自己相関

iid系列（正規ホワイトノイズなど）の累積和
ランダムウォークは「1時点前の自分の値」を含む

ランダムウォーク(RW)のシミュレーション

1日目のRW ＝ WN①

2日目のRW ＝ 1日目のRW ＋ WN②

3日目のRW ＝ 2日目のRW ＋ WN③

4日目のRW ＝ 3日目のRW ＋ WN④

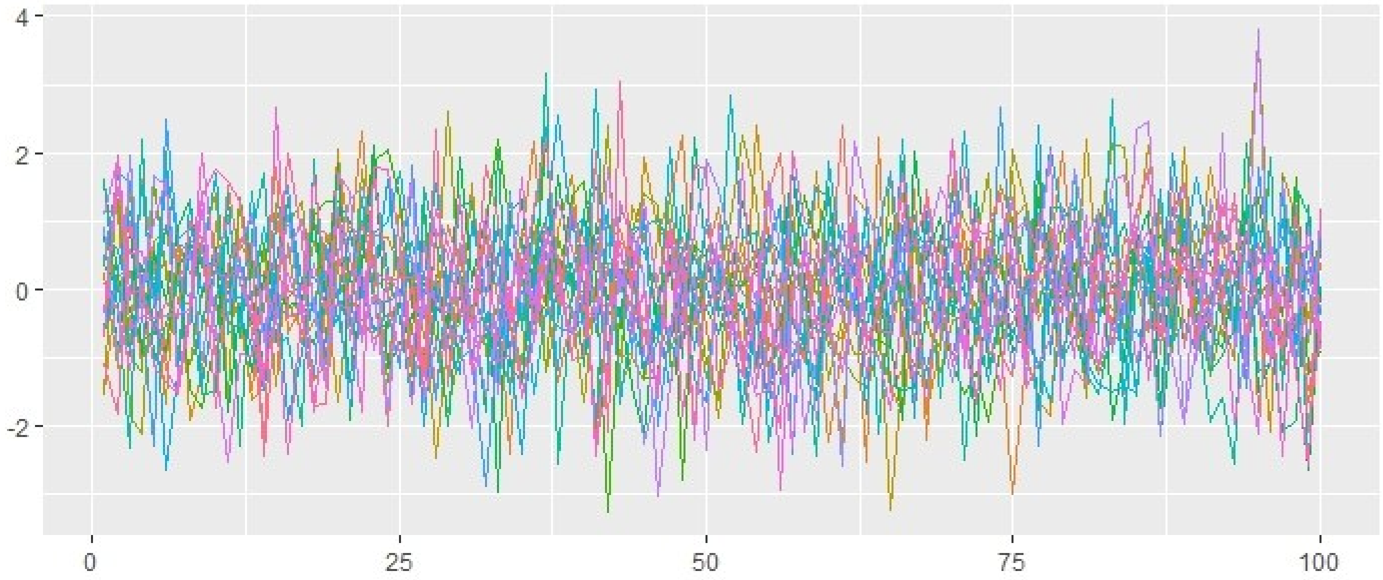
ランダムウォーク過程の初歩

ランダムウォーク系列と自己相関

iid系列（正規ホワイトノイズなど）の累積和
ランダムウォークは「1時点前の自分の値」を含む

同じ値を使う→似ている
→正の自己相関がある！

正規ホワイトノイズ：自己相関0でばらつき小さい



ランダムウォーク：正の自己相関があり、ばらつき大きい

ランダムウォーク過程の初歩

2つのホワイトノイズ(WN)系列の散布図

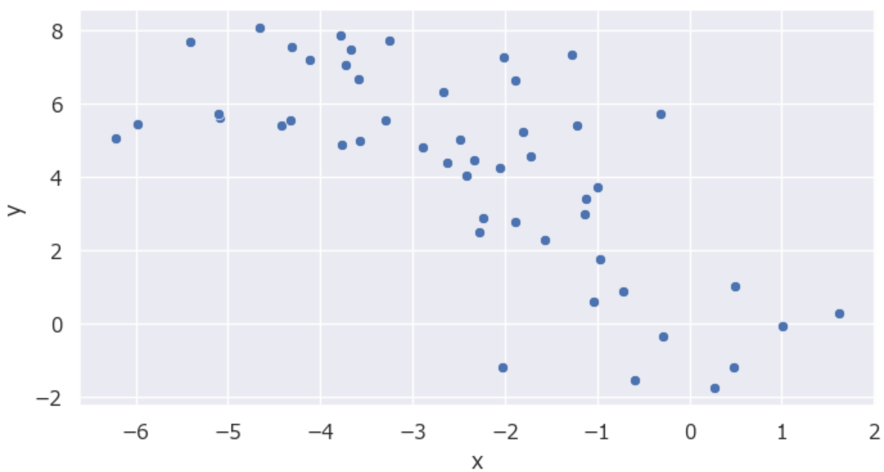
WN系列はばらつきが小さいので、
大体同じようなグラフになる

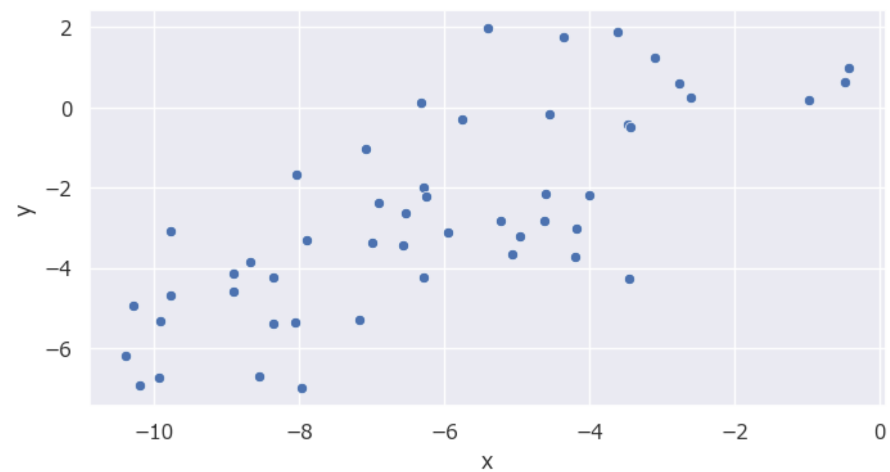
回帰分析を実行すると
傾きは大体0になる

ランダムウォーク過程の初歩

2つのランダムウォーク(RW)系列の散布図

RW系列はばらつきが大きいので、
かなりグラフの形が変わる





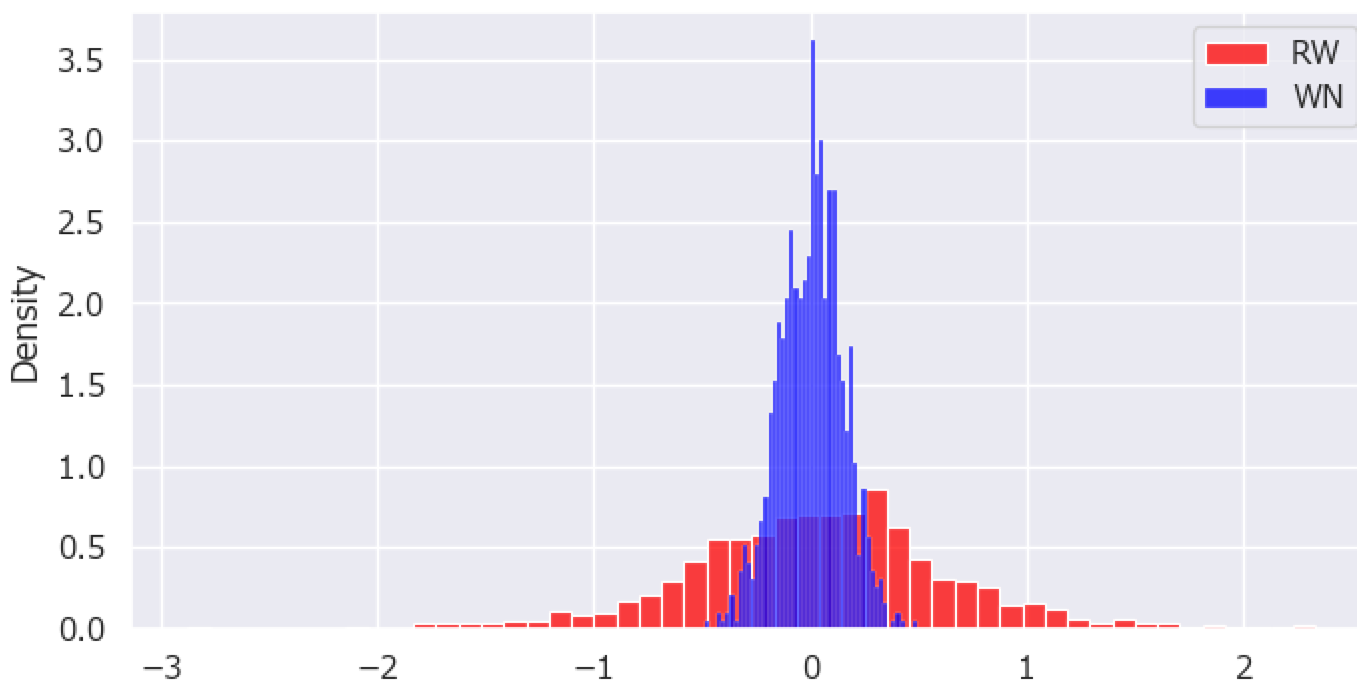
マイナスの傾き？

プラスの傾き？

ランダムウォーク過程の初歩

WN、RW系列への回帰分析

RW系列はばらつきが大きいので、
0から離れた傾きが出やすい



ランダムウォーク過程の初歩

2つのランダムウォーク(RW)系列への回帰分析

本来は何の関係もない2つのRW系列であっても、
傾きのばらつきが大きいので「たまたま大きな傾き」
等が出やすい
→本来なら何の関係もないのに
「傾きが0と有意に異なるかどうか」の検定をすると
有意差がでることが頻繁にある

RW系列に対しては、
通常の仮説検定や区間推定は使えない！

状態空間モデル(SSM)

状態空間モデル(SSM)

1. 状態空間モデルの概要

2. ローカルレベルモデル

状態空間モデルとは

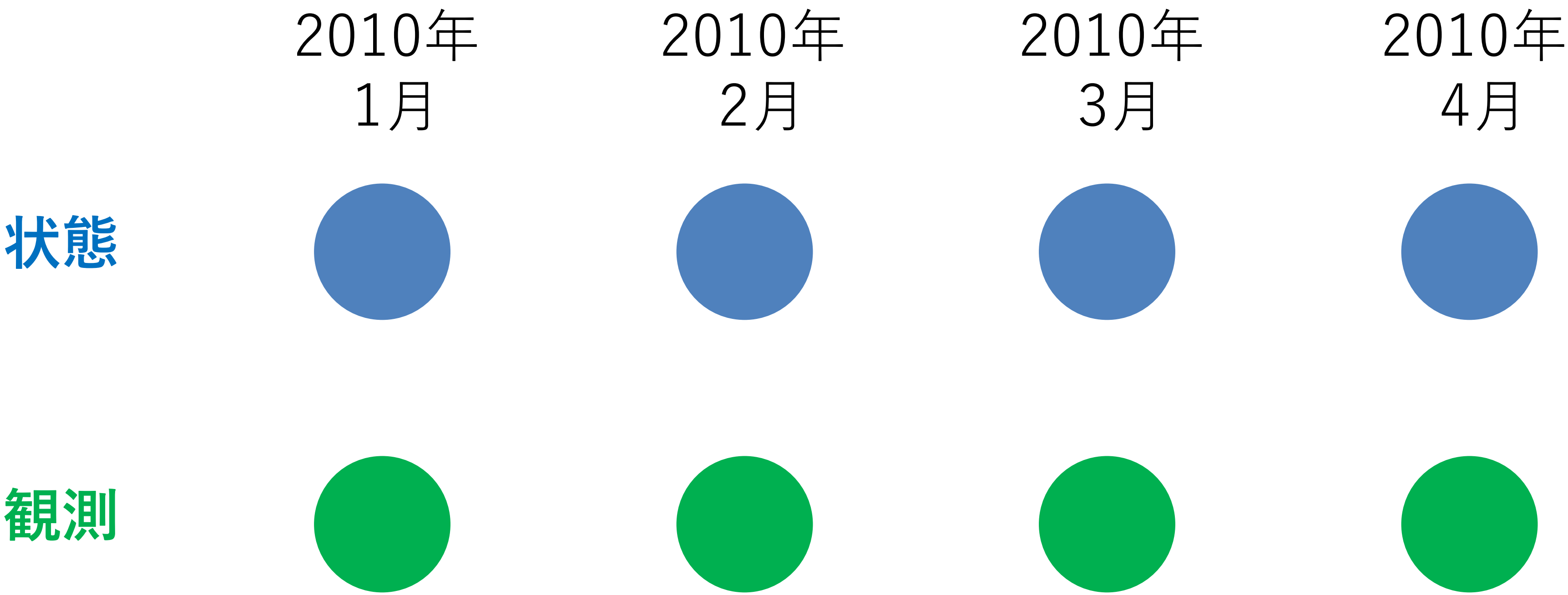
<u>状態空間モデル</u>

時系列データにも適用できる統計モデルの1種
目に見えない「状態」の存在を仮定しているのが特徴

State Space Models略してSSMと呼ぶこともある

なお、線形ガウス状態空間モデルは、動的線形モデル
Dynamic Linear Models略してDLMと呼ぶこともある
→今回のテーマはDLM

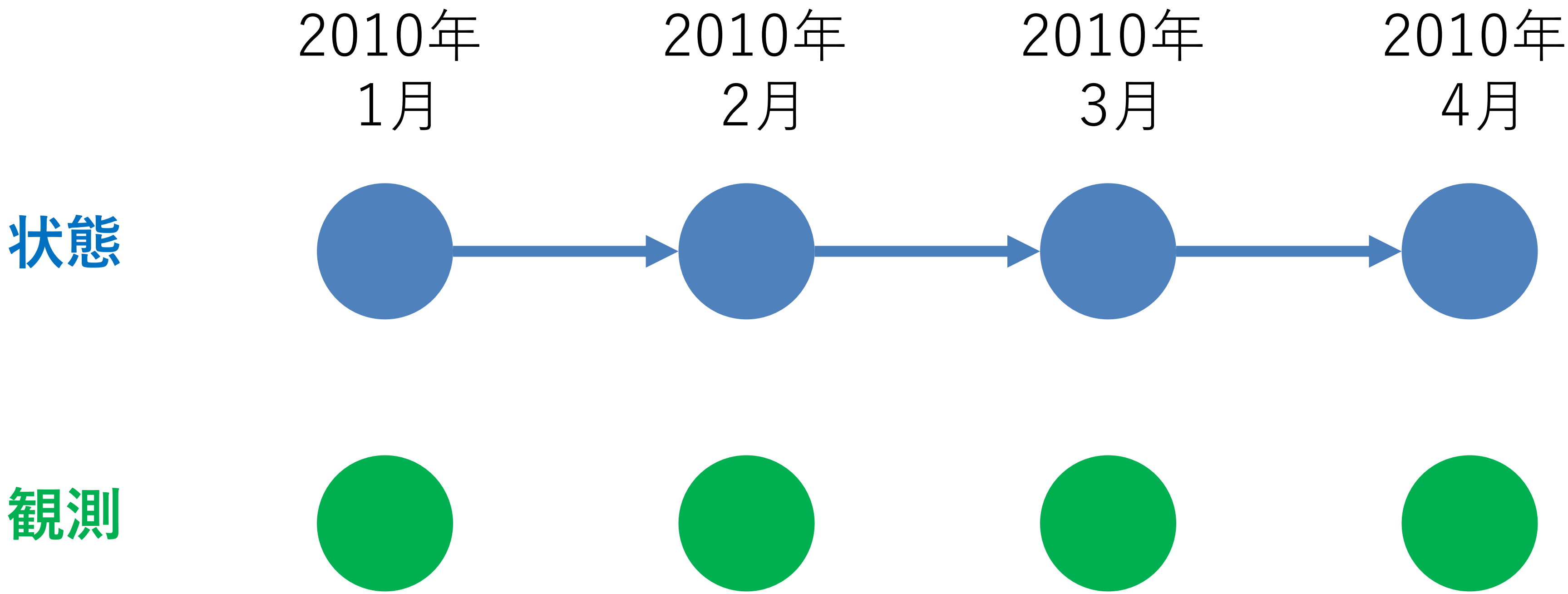
状態空間モデルとは



状態空間モデルとは

目にすることのできない状態と
目に見える観測とに分けてモデル化する

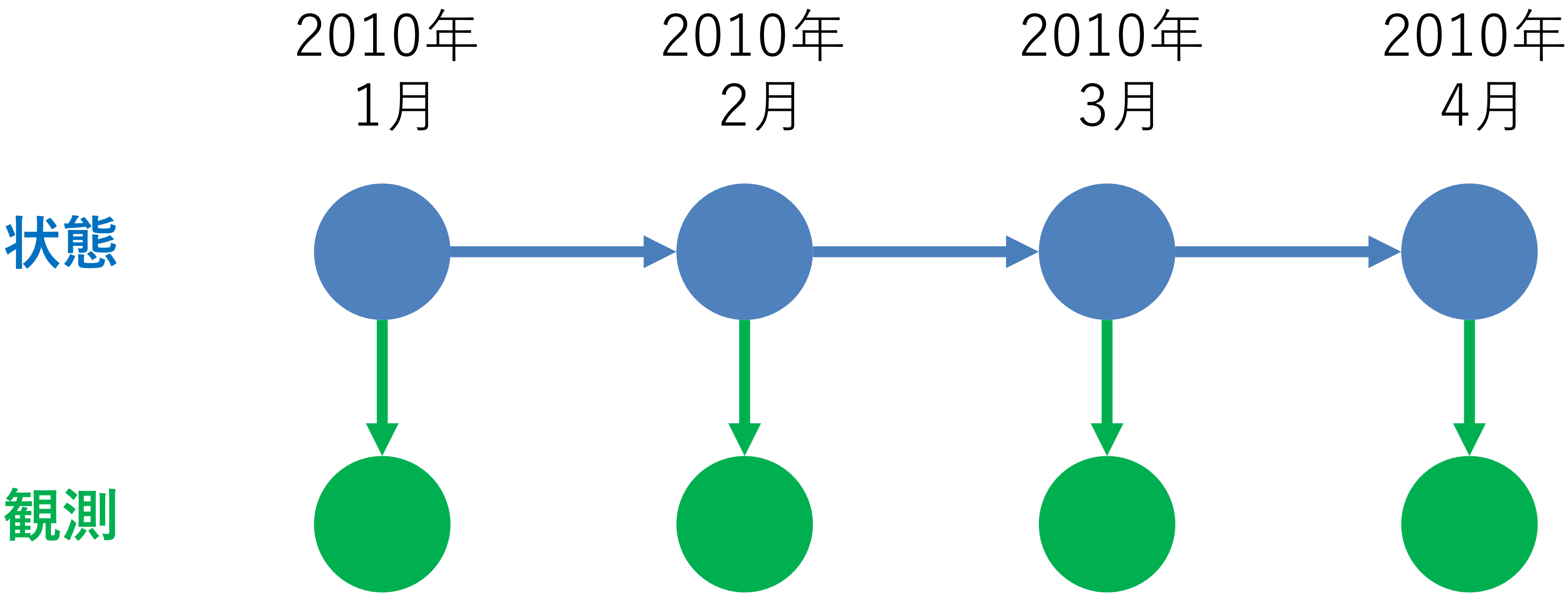
状態方程式



状態方程式とは

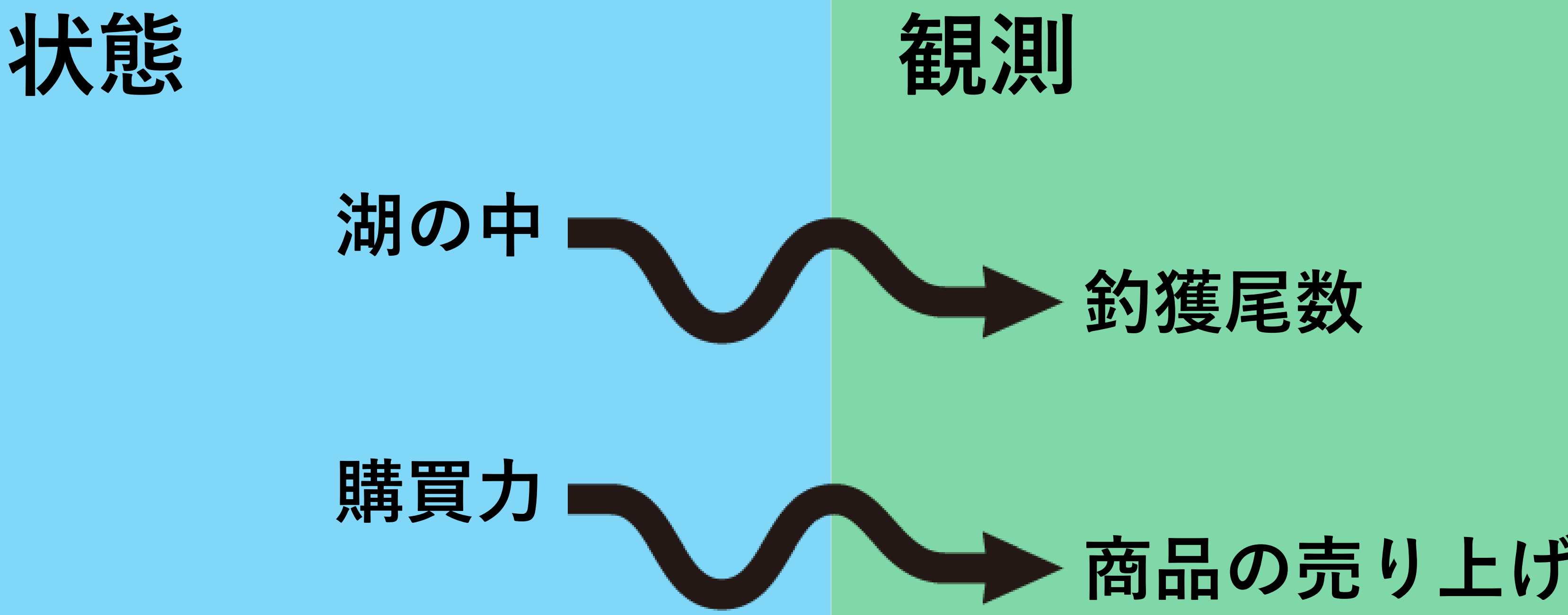
状態は過去の影響を受けている
状態方程式で過去と未来の関係を表現

観測方程式



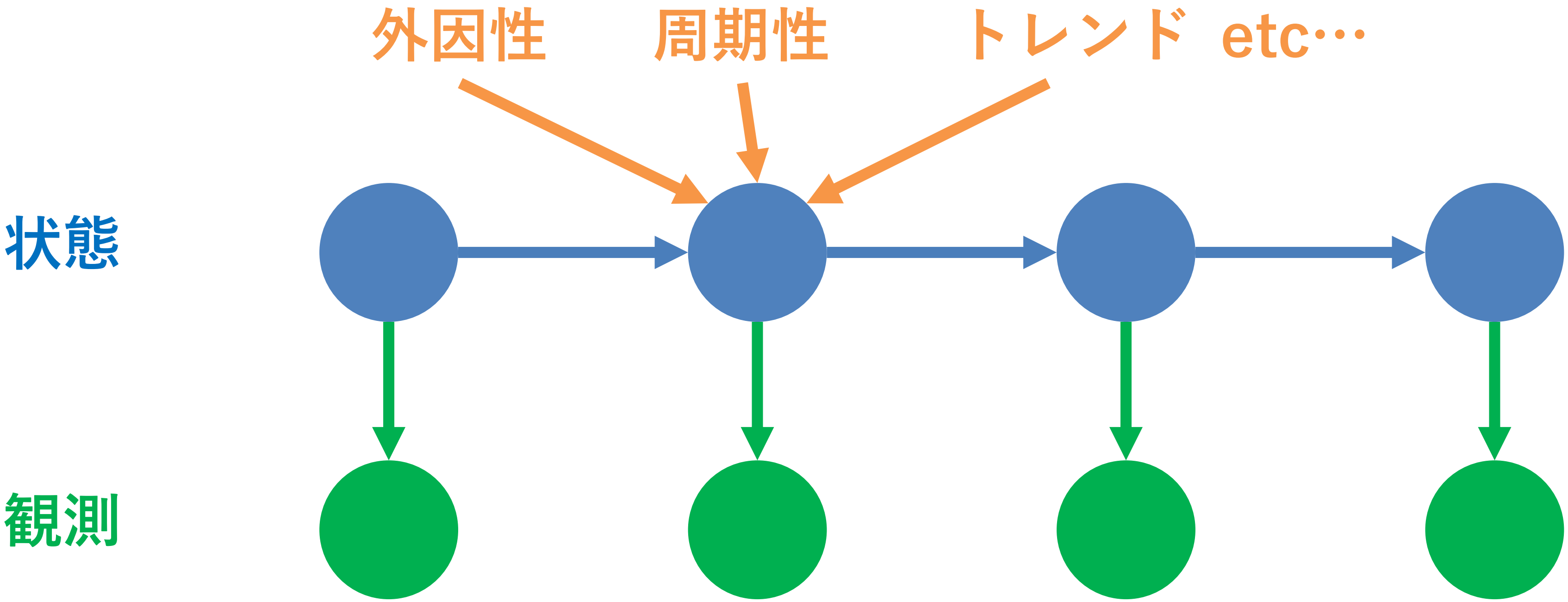
観測方程式とは

観測は現在の状態から得られる
観測方程式で観測値の得られ方を表現



なぜ目に見えない**状態**の存在を
モデルに組み込むのだろうか？

状態を仮定する意義



複雑な状態の変化を柔軟にモデリング
周期やトレンドの成分を個別に分解できる

状態空間モデル(SSM)基礎

1. 状態空間モデルの概要
2. ローカルレベルモデル

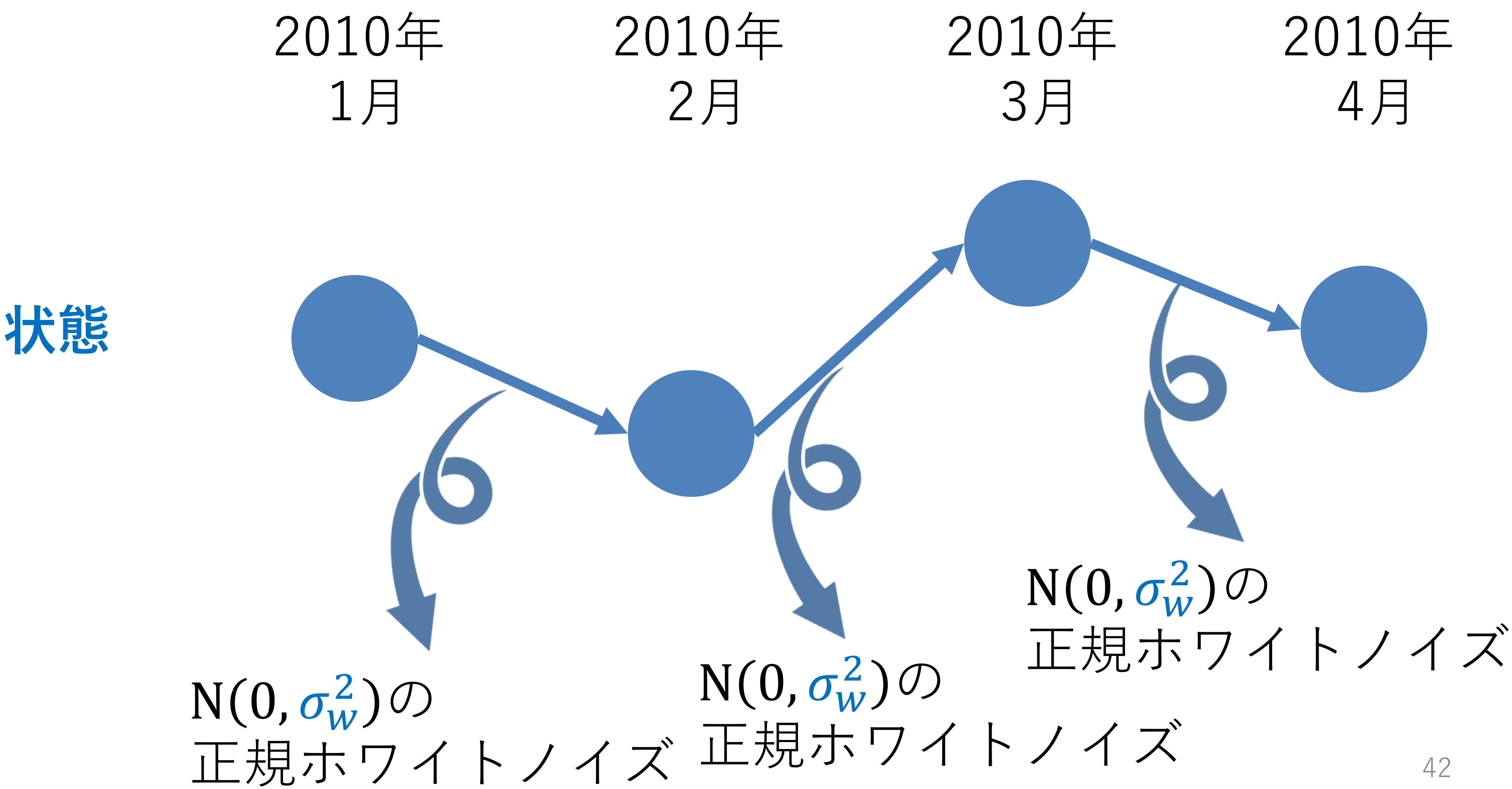
ローカルレベルモデルとは

状態空間モデルの「型」の1つ。とてもシンプル

別名は
「ランダムウォーク＋ノイズ」モデル

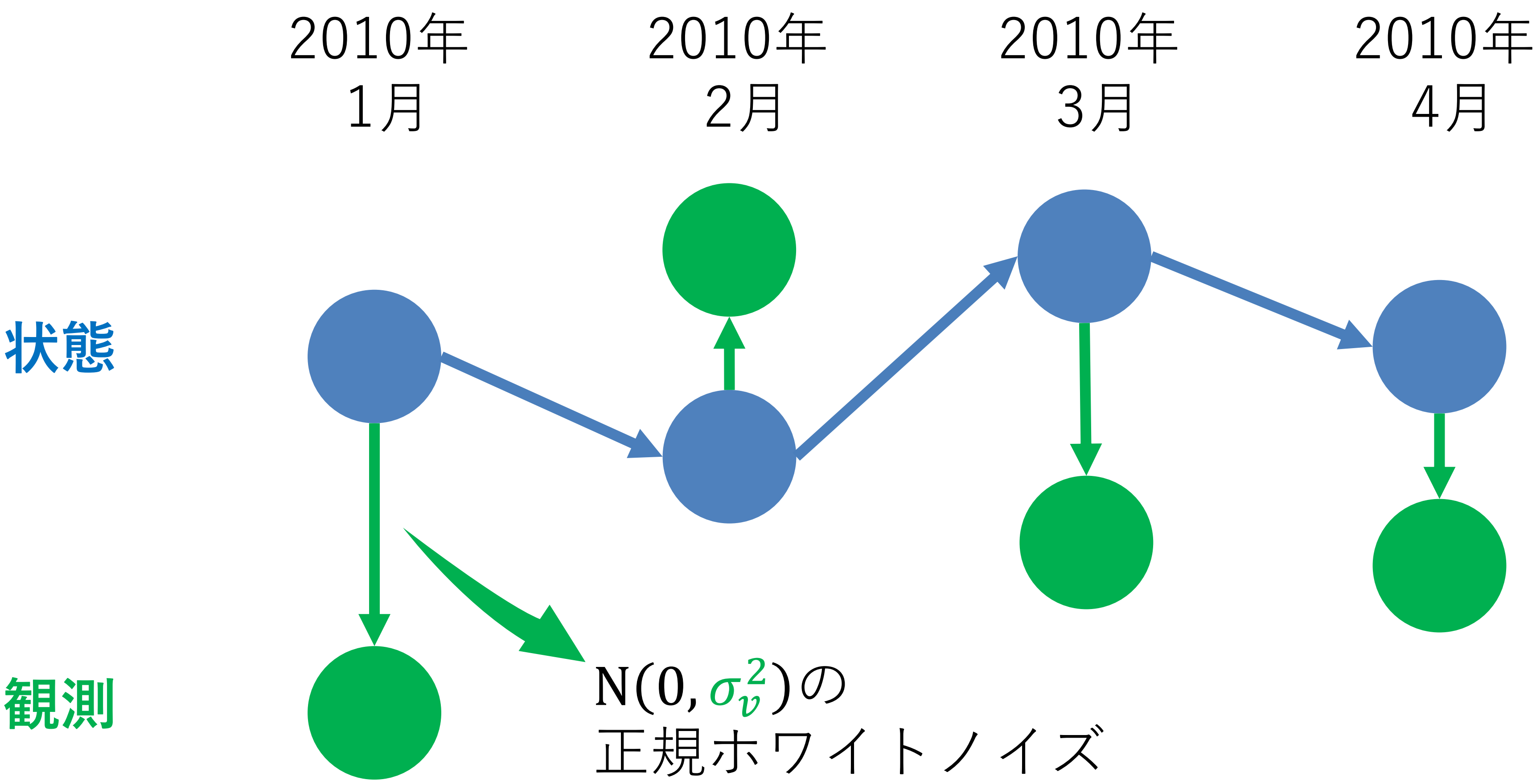
ローカルレベルモデルとは

状態はランダムウォーク的に変化(正規ホワイトノイズの累積和)



ローカルレベルモデルとは

状態にノイズが加わって観測が得られる



用語の整理

ランダムウォーク

iid系列の累積和（非定常）
iid系列として正規ホワイトノイズを今回は仮定

過程誤差

状態が「次の時点の状態」に遷移する際のノイズのこと
過程誤差の大きさは、分散σ_w^2で表現
過程誤差が大きいと、状態が大きく変化しやすい

観測誤差

状態から観測が得られる際に加わるノイズのこと
観測誤差の大きさは、分散σ_v^2で表現

ローカルレベルモデルの構造

状態方程式 $\mu_t = \mu_{t-1} + w_t, \quad w_t \sim \mathrm{N}(0, \sigma_w^2)$

今期の状態 ＝ 前期の状態 ＋ 過程誤差

ローカルレベルモデルの構造

状態方程式　$\mu_t = \mu_{t-1} + w_t,\quad w_t \sim \mathrm{N}(0, \sigma_w^2)$

観測方程式　$y_t = \mu_t + v_t,\quad v_t \sim \mathrm{N}(0, \sigma_v^2)$

今期の観測 ＝ 今期の状態 ＋ 観測誤差

ローカルレベルモデルの構造

状態方程式	今期の状態	=	前期の状態	+	過程誤差
観測方程式	今期の観測	=	今期の状態	+	観測誤差

状態空間モデルの推定方法の概要

内容

1. 状態空間モデルを推定する流れ
2. カルマンフィルタ
3. 平滑化
4. 状態空間モデルにおける最尤法

状態空間モデルの推定

状態空間モデルには複数の推定方法がある

線形ガウス状態空間モデルの推定
非線形非ガウス状態空間モデルの推定
→異なる手法となる

線形ガウス状態空間モデルの推定方法

カルマンフィルタと最尤法の組み合わせがおすすめ
→計算速度が速い！

状態空間モデルを推定する流れ

Step1　　カルマンフィルタによる状態推定

状態の推定のためには、パラメータが必要
過程誤差の分散（ σ_w^2 ）
観測誤差の分散（ σ_v^2 ）

Step1の段階では
暫定的なパラメータを使って状態を推定するしかない

状態空間モデルを推定する流れ

Step1　カルマンフィルタによる状態推定

Step2　最尤法によるパラメータの推定

「暫定的なパラメータ」を使って状態を推定したのち、
試行錯誤的にパラメータを修正していくイメージ

状態空間モデルを推定する流れ

Step1　　カルマンフィルタによる状態推定

Step2　　最尤法によるパラメータの推定

Step3　　カルマンフィルタによる状態の再推定
　　　　　平滑化による、状態推定

平滑化はカルマンフィルタと違って、
データ全部を余すことなく使って状態を推定する

状態空間モデルを推定する流れ

Step1　カルマンフィルタによる状態推定　**状態**

Step2　最尤法によるパラメータの推定　**パラメータ**

Step3　カルマンフィルタによる状態の再推定　**状態**
平滑化による、状態推定

今回は先に**状態**の推定方法をまとめて説明
最後に**パラメータ**推定の方法を説明

内容

1. 状態空間モデルを推定する流れ
2. カルマンフィルタ
3. 平滑化
4. 状態空間モデルにおける最尤法

状態推定

パラメータ推定

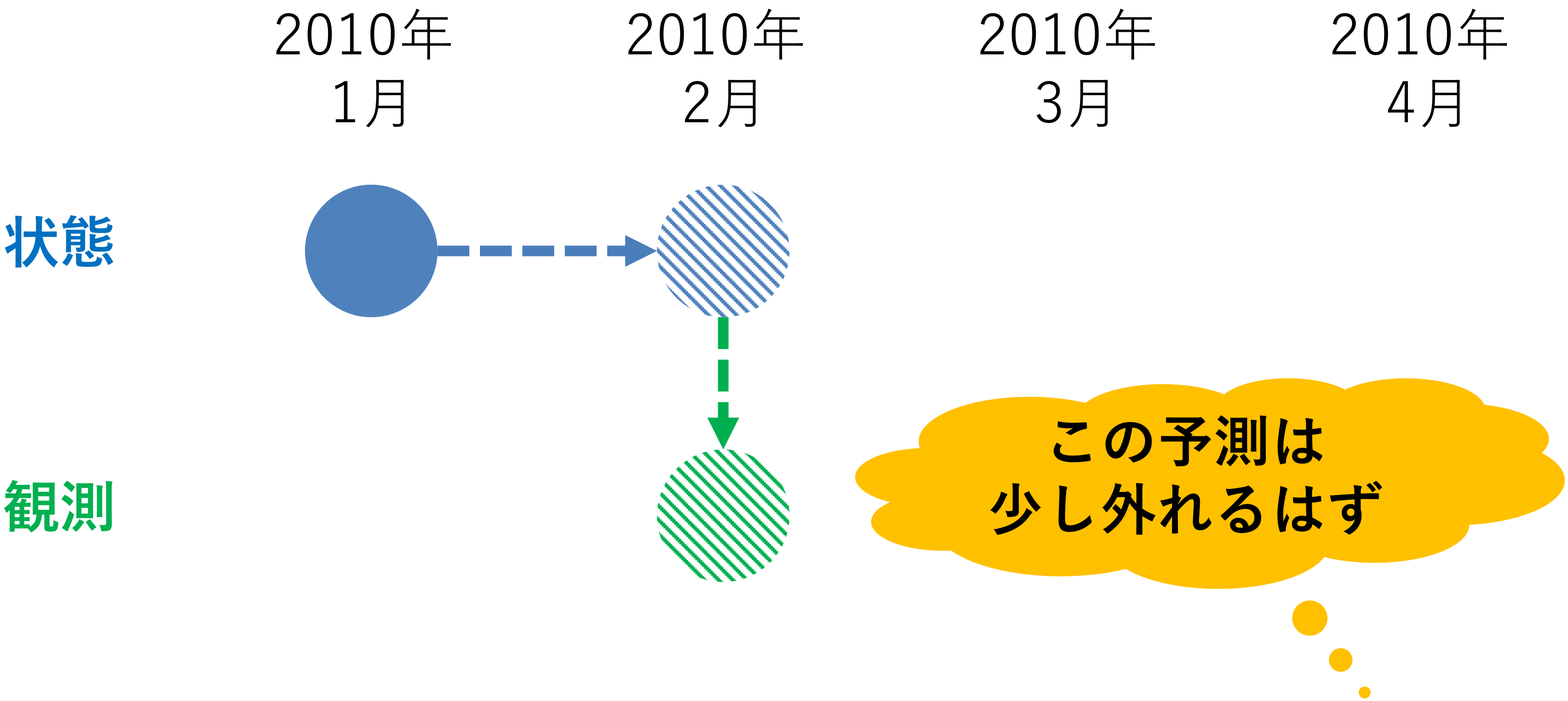
カルマンフィルタとは

カルマンフィルタ

「予測→フィルタリング」の繰り返し計算により
効率的に状態を推定する計算手法

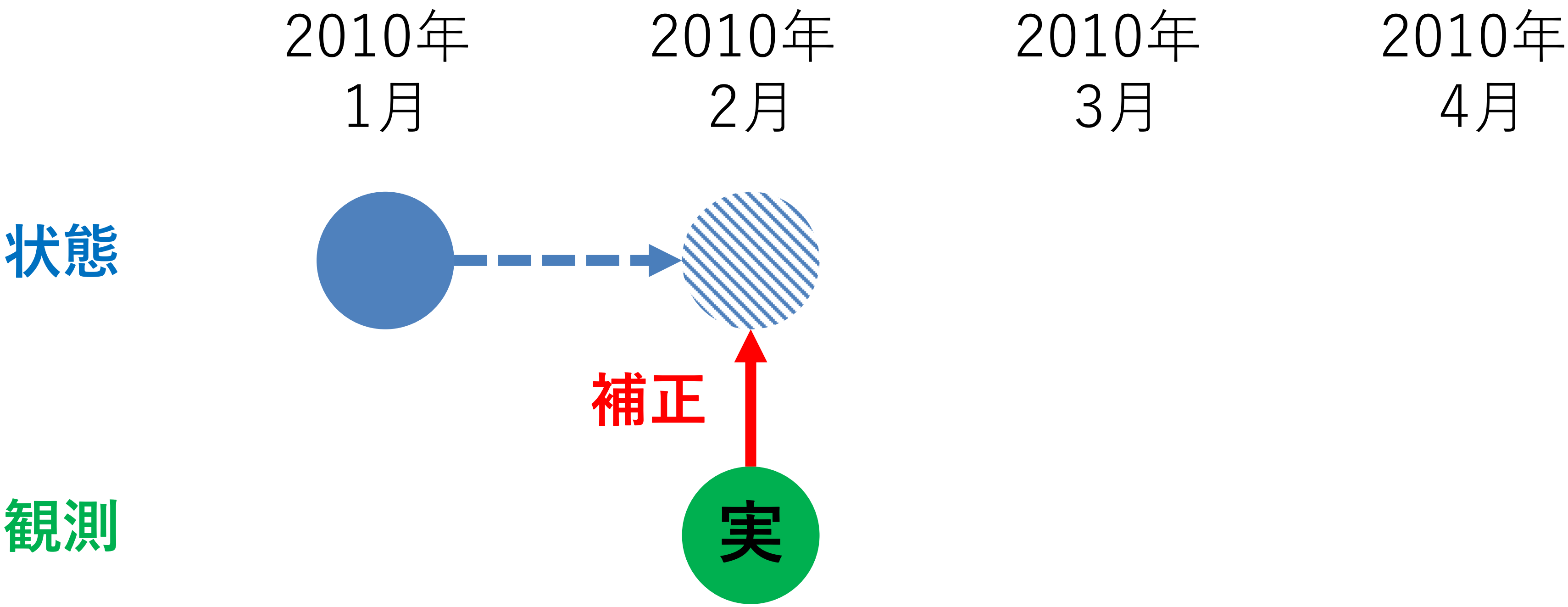
まずは「予測」と「フィルタリング」の
イメージをつかんでもらいます

カルマンフィルタの流れ：予測



前期の状態から今期の状態を予測
今期の状態から今期の観測を予測

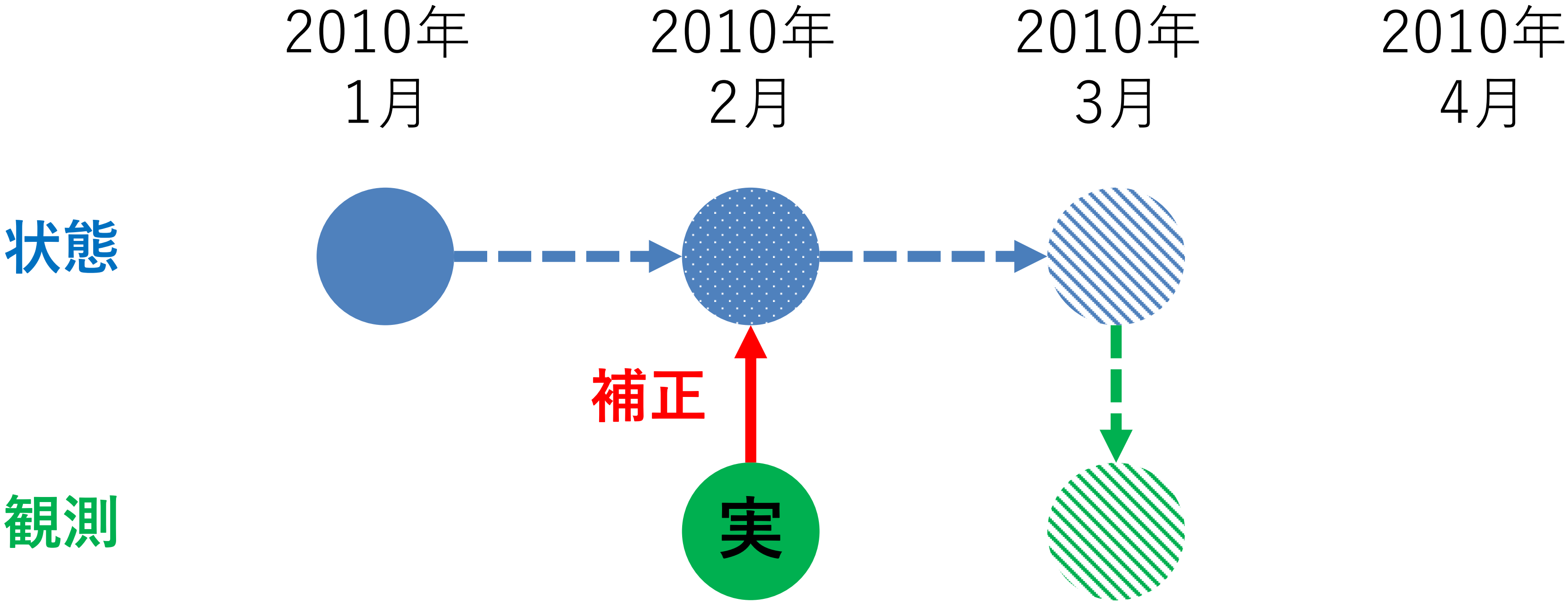
カルマンフィルタの流れ：フィルタリング



フィルタリングとは

今期の「実際に得られた実測値」を使い予測された今期の状態を補正すること

カルマンフィルタの流れ：予測



カルマンフィルタの流れ

「補正された状態」に基づき
次回の予測を行う

用語の整理

<u>フィルタリング</u>

実測値に基づいて、状態を補正すること
「今期の実測値」を使って「今期の状態」を補正

<u>フィルタ化推定量</u>

フィルタリングによって補正された状態の値

カルマンフィルタ

フィルタリングは「状態の補正」をするもの

↓

どのように「補正」をするのが正しい？
→これから「補正」の方法論の説明に移ります

カルマンフィルタ

案1：実測値と同じ値になるまで状態を補正

実測値を完全に信用
補正後は、観測と状態が等しくなる

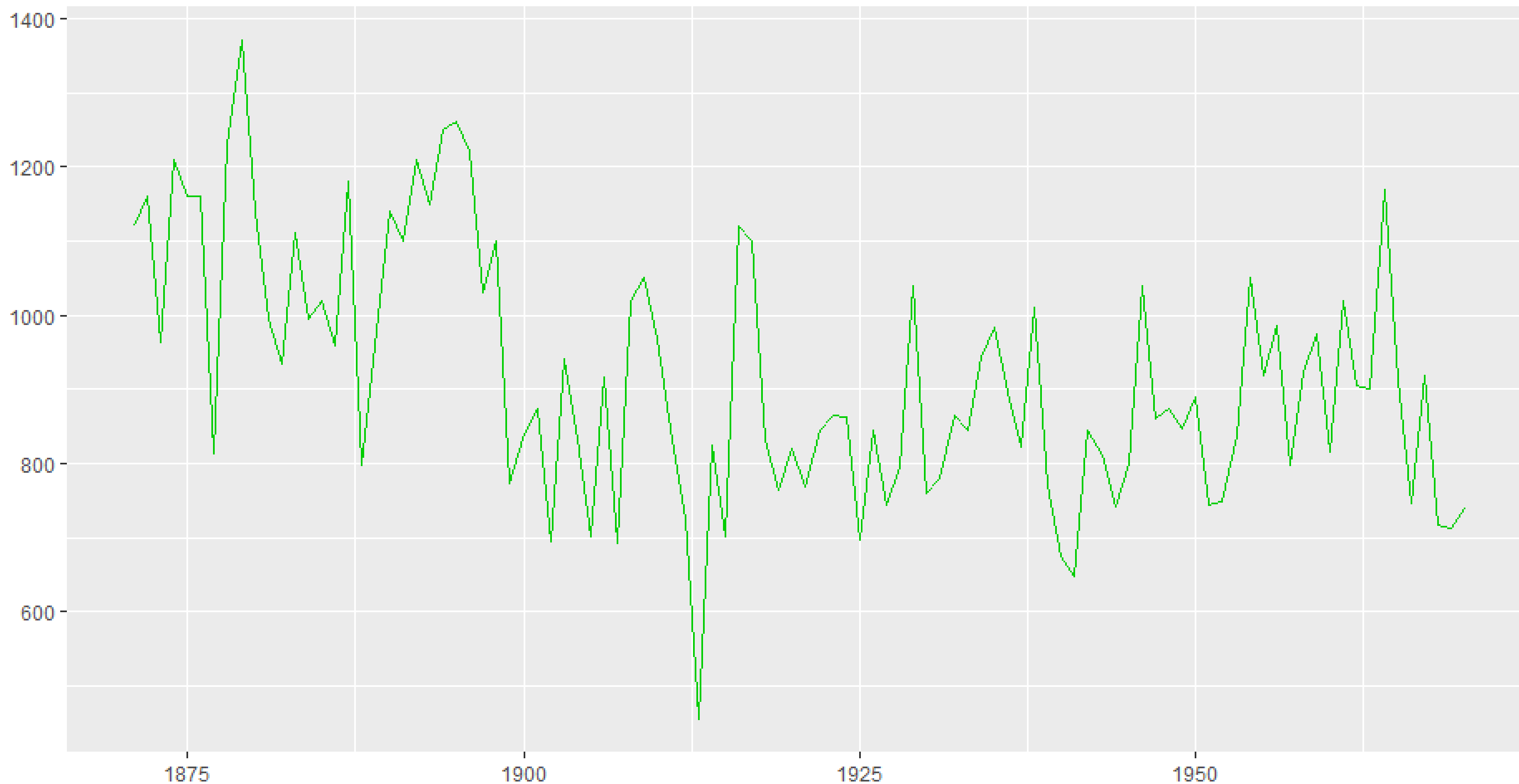
案2：実測値は無視して状態を補正しない

実測値を完全に無視
観測による状態の補正が行われない

ちょうどよい塩梅で補正したい

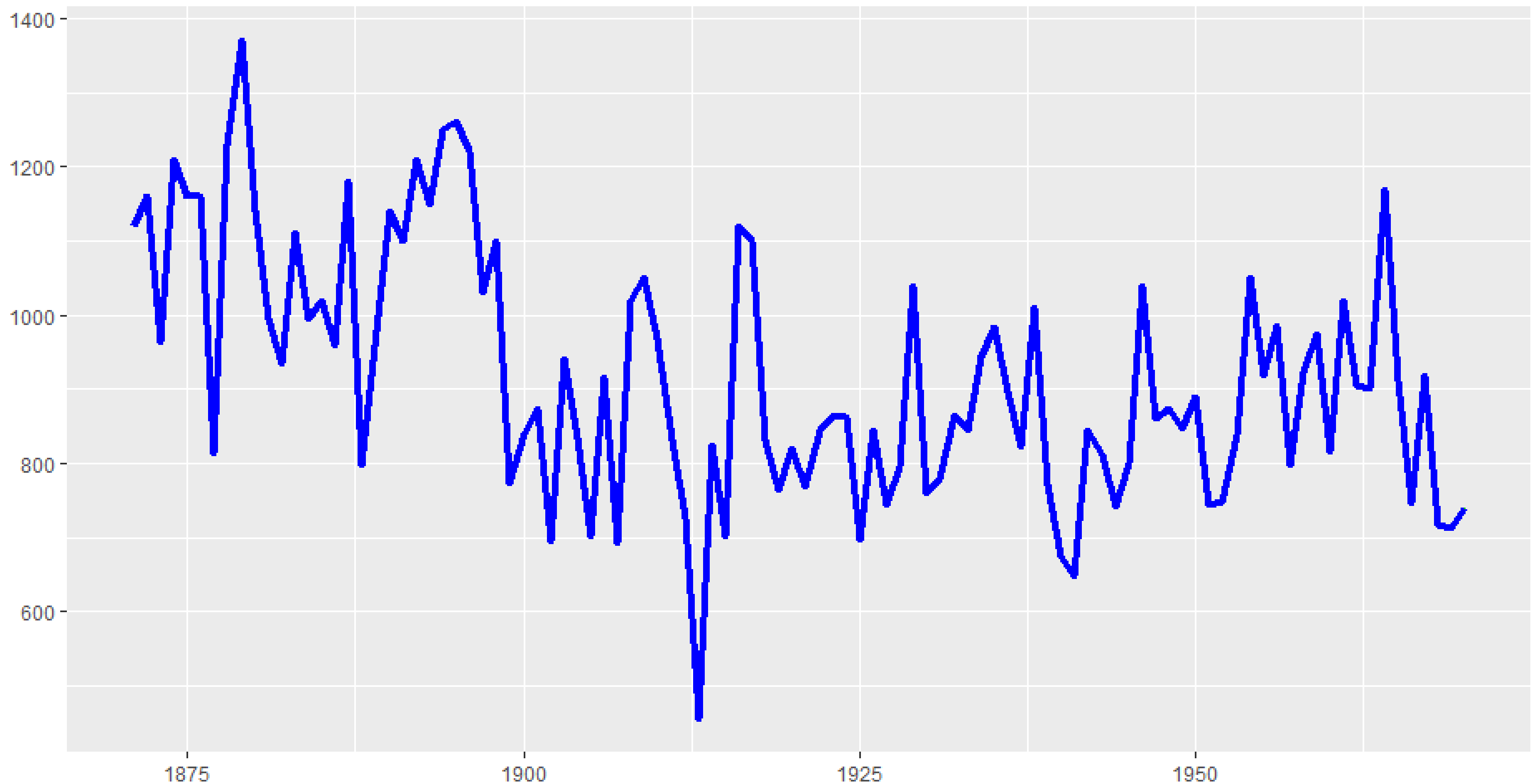
カルマンフィルタ

ナイル川の流量データ（観測値）



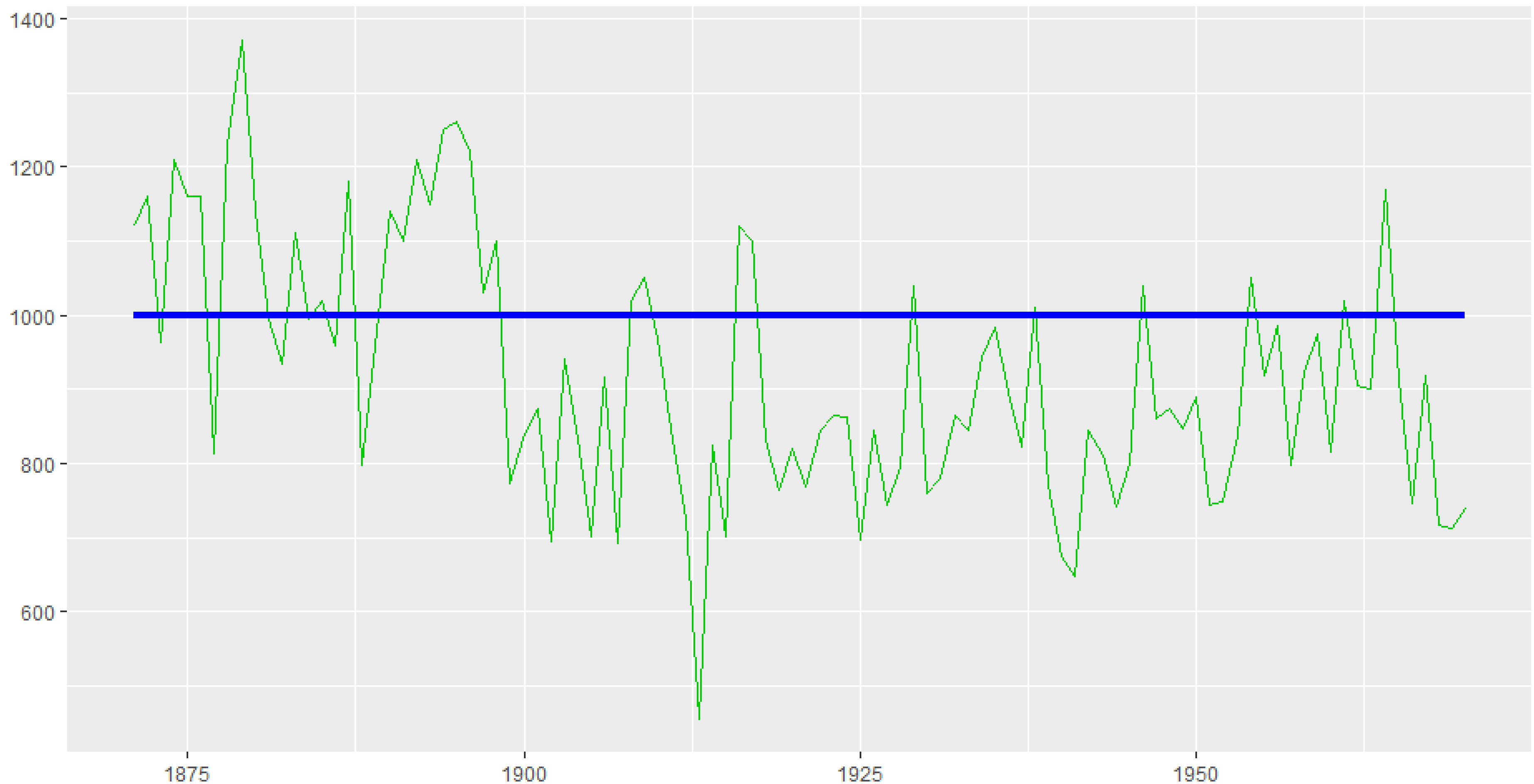
カルマンフィルタ

案1：実測値と同じ値になるまで状態を補正



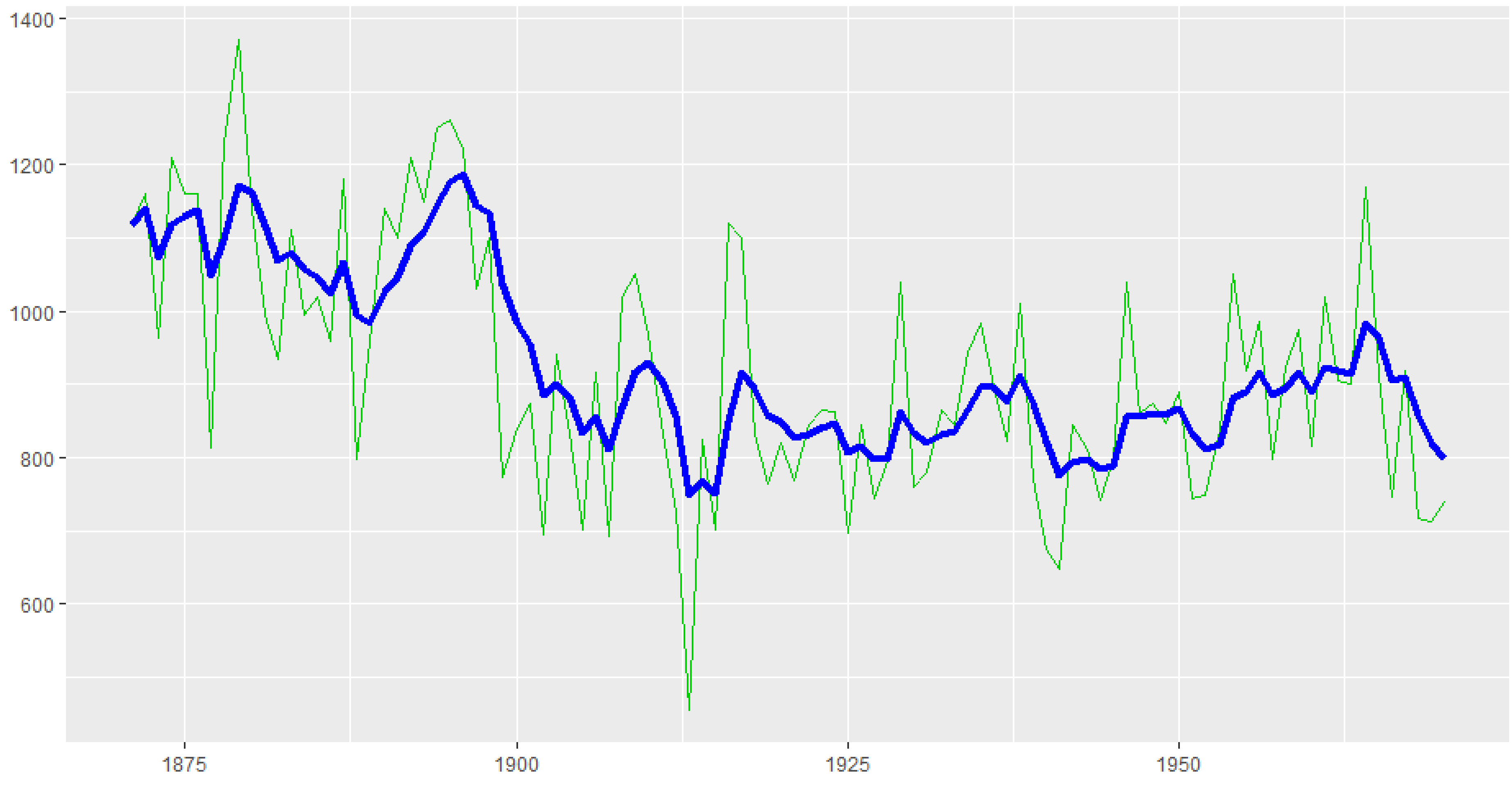
カルマンフィルタ

案2：実測値は無視して状態を補正しない



カルマンフィルタ

ちょうどいい塩梅



カルマンゲイン

フィルタ化推定量（補正後の状態）
＝補正前の状態＋カルマンゲイン×予測残差

実測値－予測された観測

ゲインが1なら：実測値と同じ値になるまで状態を補正

ゲインが0なら：実測値は無視して状態を補正しない

※　カルマンゲインを推定する詳細は略

カルマンフィルタの流れ

Step1：前期の情報から今期を予測する

Step2：フィルタリングして状態を補正

Step3：補正後の状態からさらに予測……

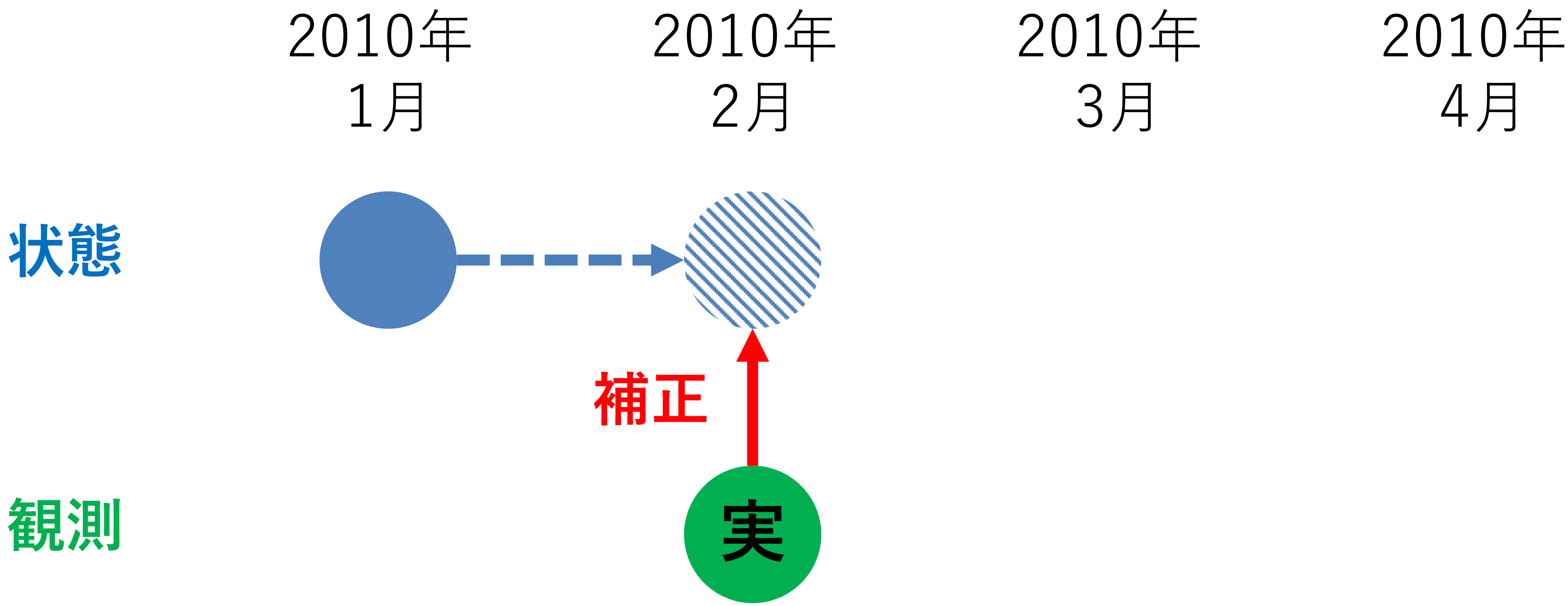
内容

1. 状態空間モデルを推定する流れ
2. カルマンフィルタ
3. 平滑化
4. 状態空間モデルにおける最尤法

状態推定

パラメータ推定

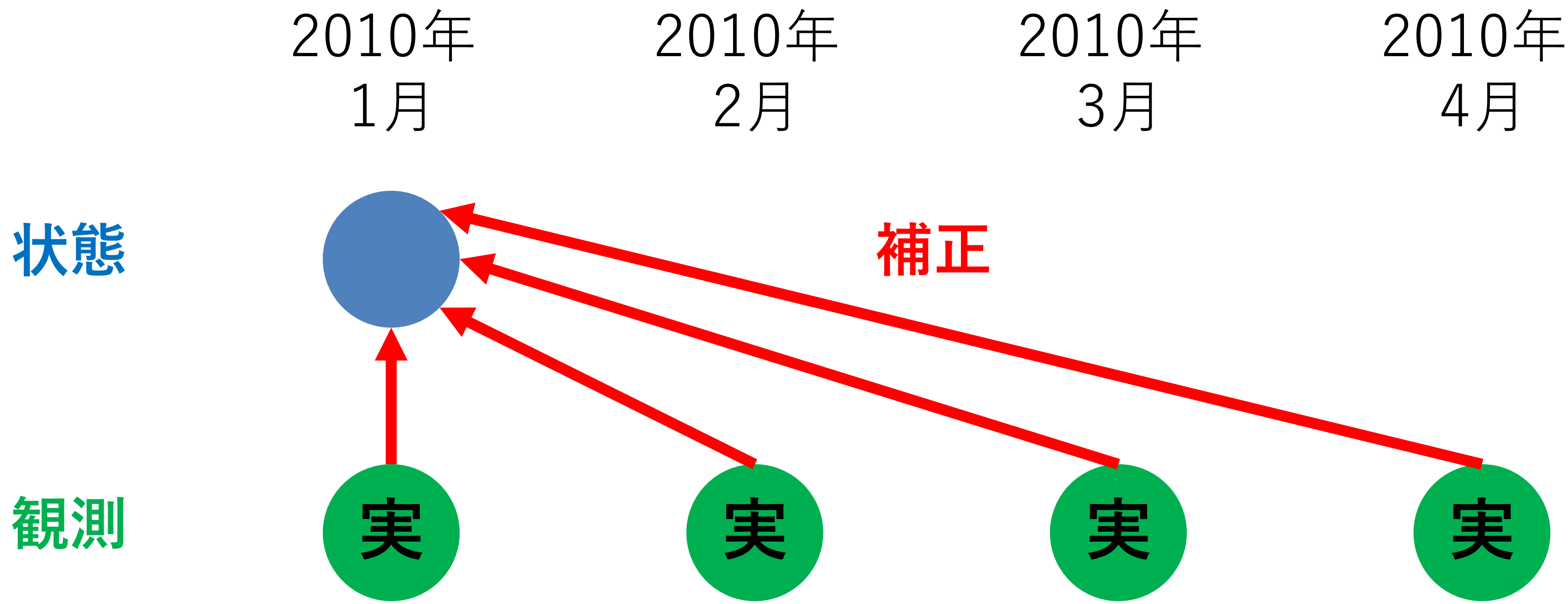
フィルタリングの復習



フィルタリングとは

今期の「実際に得られた実測値」を使い
今期の「状態」を補正する

平滑化



平滑化とは

全ての「実際に得られた実測値」を使い古い時点の「状態」を補正する

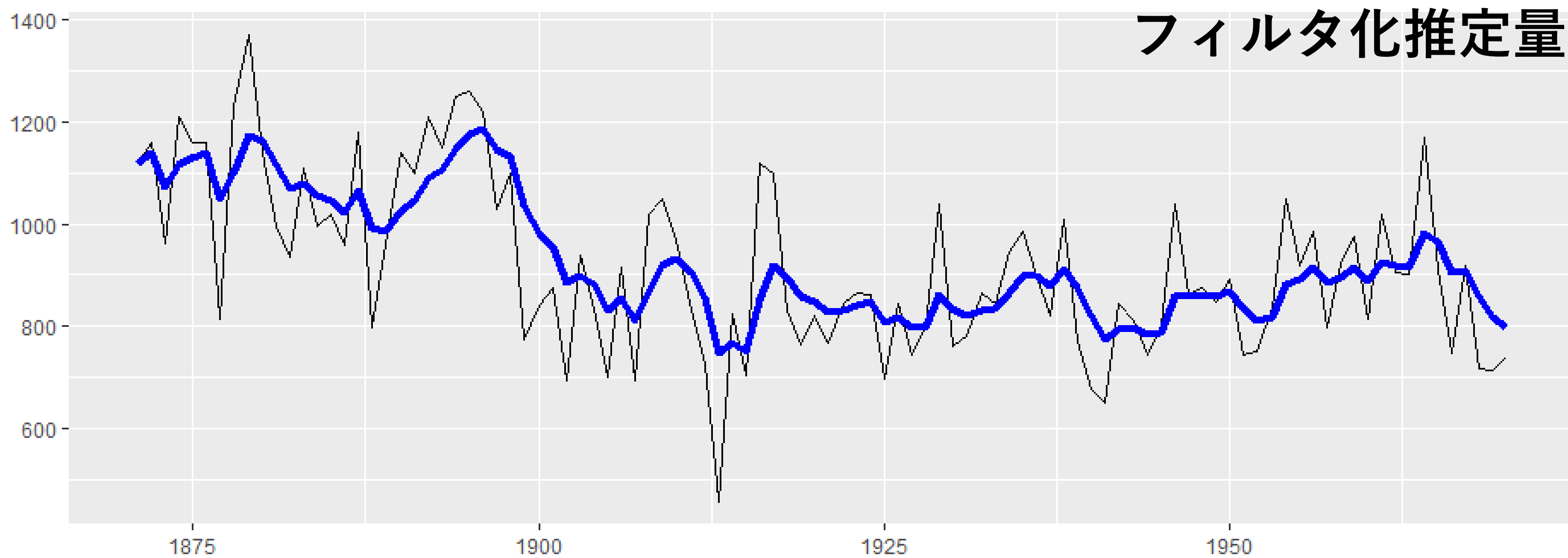
平滑化の補足

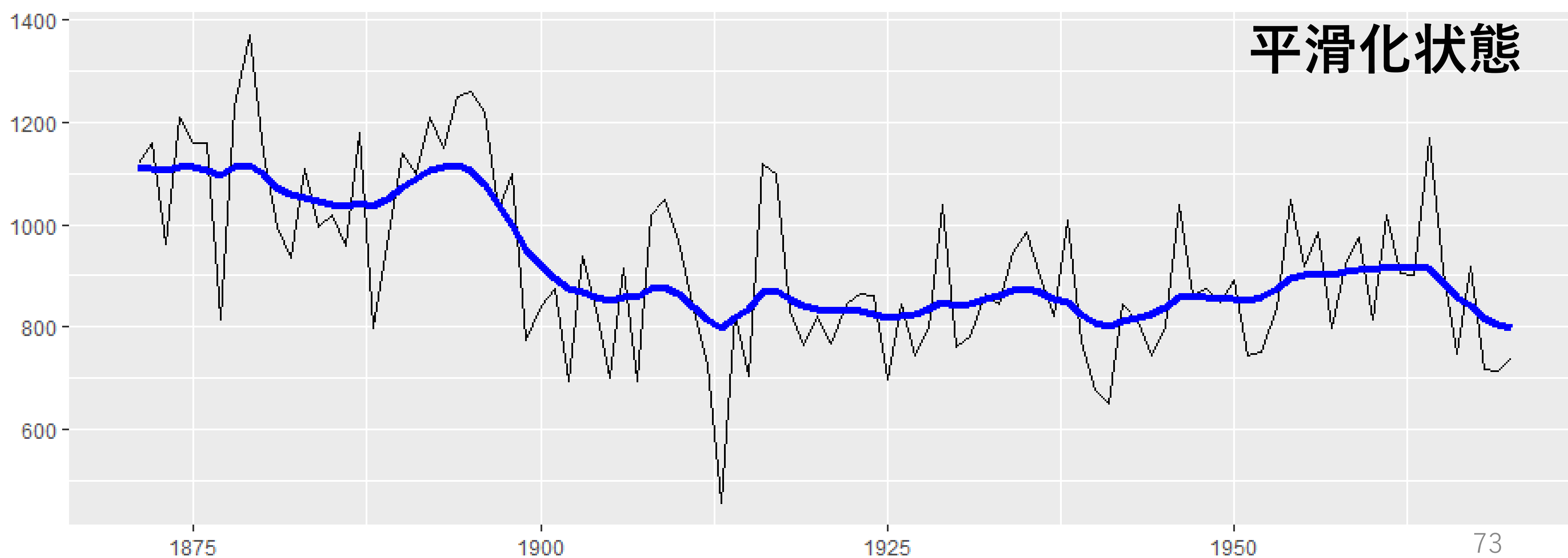
<u>平滑化状態</u>

平滑化を用いて推定された状態のこと

最新時点においては（未来のデータがないので）、
フィルタ化推定量と平滑化状態は一致する

平滑化状態の方が「滑らか」になる


フィルタ化推定量
1400
1200
1000
800
600
1875
1900
1925
1950


平滑化状態
1400
1200
1000
800
600
1875
1900
1925
1950

内容

1. 状態空間モデルを推定する流れ
2. カルマンフィルタ
3. 平滑化
4. 状態空間モデルにおける最尤法

状態推定

パラメータ推定

カルマンフィルタにおける尤度

データ：　観測値の予測残差

平均値が0
正規分布に従う

パラメータ：過程誤差の分散（ σ_w^2 ）
観測誤差の分散（ σ_v^2 ）

尤度：　$\prod$ N(0, 観測値の予測誤差の分散)

前期の状態の予測誤差の分散 + σ_w^2 + σ_v^2

最尤法まとめ

Step1　暫定的なパラメータ（σ_w^2, σ_v^2）を指定して
カルマンフィルタを実行する

Step2　想定される「予測誤差の大きさ」と
実際の予測残差を比較する

↓

尤度が増えるように、パラメータを微調整する

Step3　Step2を繰り返して、
最も尤度が高くなるパラメータを採用する

「慣れ」のモデル化

↓

お客様が、マーケティング施策に
慣れてしまうことを想定して分析を行おう！

「慣れ」のモデル化

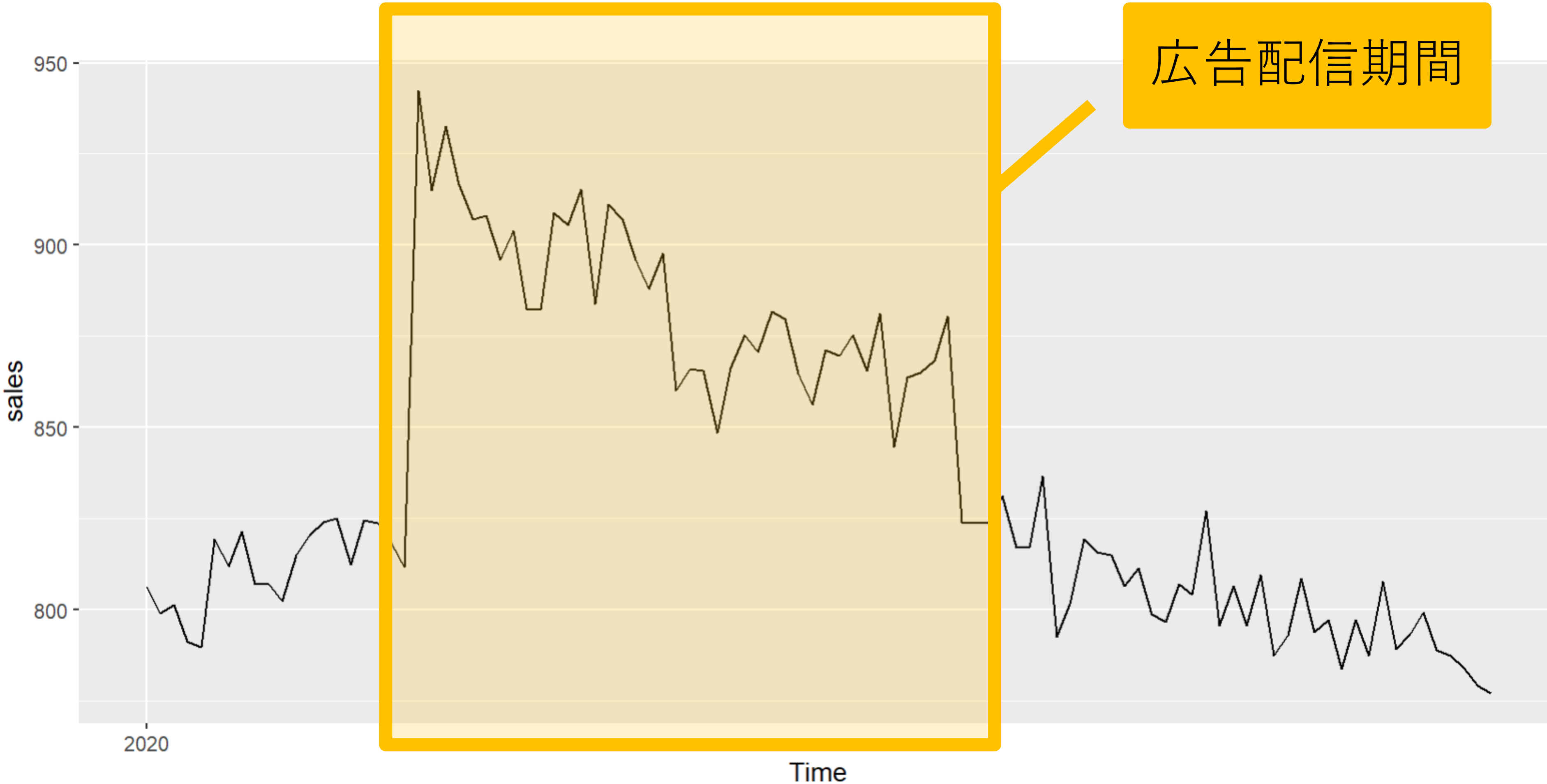
広告を配信したら、売り上げが増えたぞ！

でも、長く広告を続けていたら、
少しずつ効果が薄くなっていく……

お客様の行動が時間に応じて変化する場合は
時系列分析を行うのが良いよ

「慣れ」のモデル化

広告の効果は、長く続かないことも・・・・・・



「慣れ」のモデル化

「慣れ」をモデル化するための技術

時間に応じて、変化するパラメータを推定する

（通常の）線形回帰分析

$$y_i = Intercept + beta \cdot x_i + \varepsilon_i$$
$$\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$$

時間によってパラメータが変化する線形回帰分析

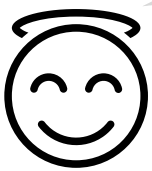
$$y_t = Intercept_t + beta_t \cdot x_t + \varepsilon_t$$
$$\varepsilon_t \sim \mathrm{Normal}(0, \sigma^2)$$

※ tはtimeを表す時間の添え字
切片も傾きも、時間によって変化する

「慣れ」のモデル化

広告フラグ＝０なら広告なし。１なら広告あり

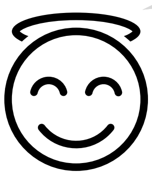
広告を出すと、売り上げは平均100万円UPだ！



売り上げ平均 = 切片 + 100 × 広告フラグ

1か月後……

広告を出しても、売り上げは40万円しか増えない！



売り上げ平均 = 切片 + 40 × 広告フラグ

「慣れ」のモデル化

広告を配信するのに60万円かかります

広告効果が60万円以下になるなら、
広告は出したくないな

状態空間モデルを使って
「慣れ」をモデル化しよう

内容

1. 時変係数モデルの理論

2. 時変係数モデルの推定

ローカルレベルモデルの構造（復習）

今期の状態 ＝ 前期の状態 ＋ 過程誤差

今期の観測 ＝ 今期の状態 ＋ 観測誤差

外因性を組み込む

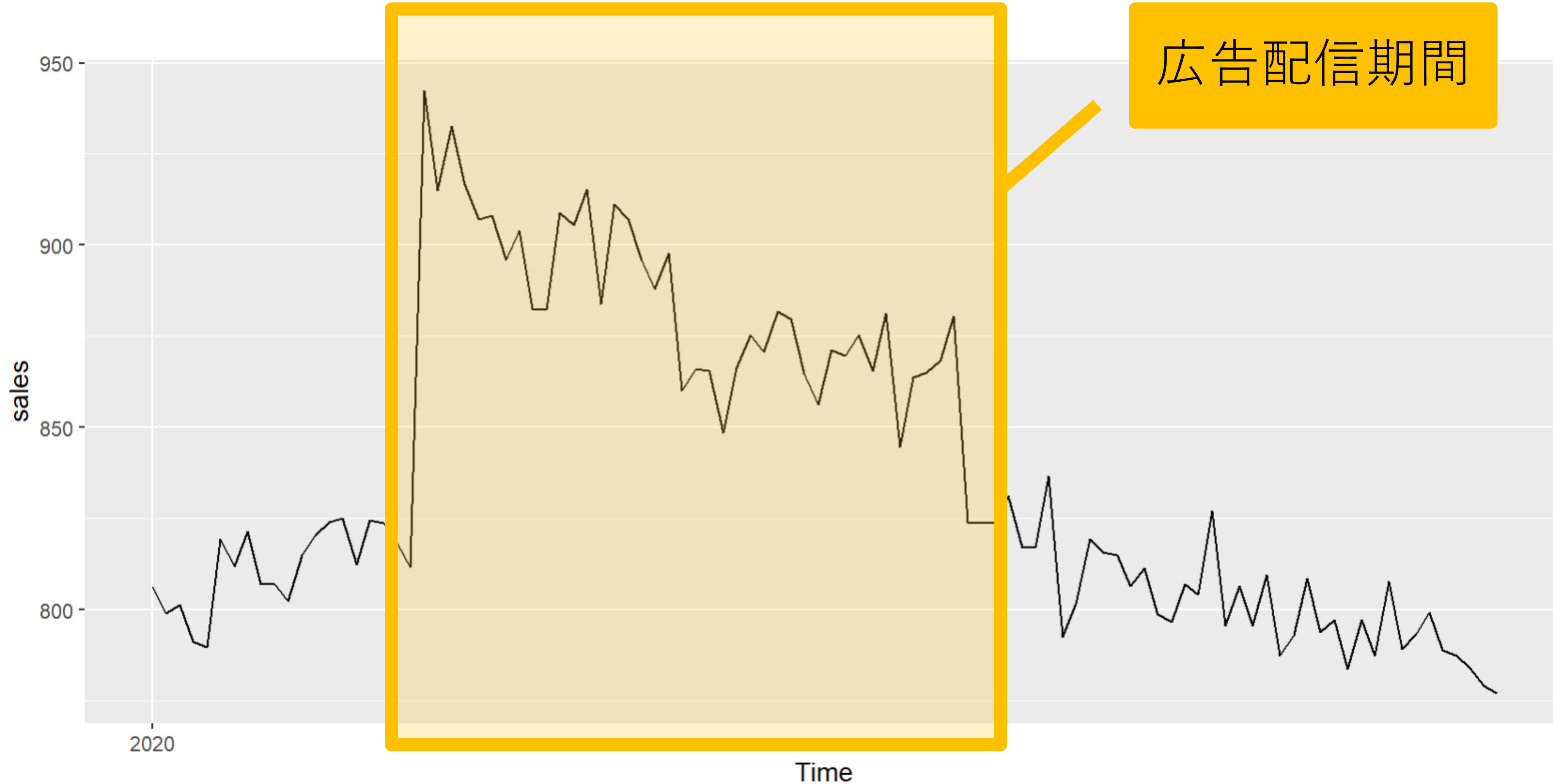
今期の状態 ＝ 前期の状態 ＋ 過程誤差

CMなどの広告のフラグ
0：広告なし／1：広告あり

今期の観測 ＝ 今期の状態 ＋ 広告効果 ＋ 観測誤差

時変係数モデル

広告の効果は、長く続かないことも・・・・・・



時変係数モデル

広告を出すと売り上げが●万円増える

↓

広告の効果を表すパラメータが、
ランダムウォークに従って変化するとみなす

↓

最初は「広告があると100万円UP」だったのが
1月後には、「40万円UP」まで下がっている、等

時変係数モデル

今期の状態 ＝ 今期の水準 ＋ 今期の係数 × 広告フラグ

水準（レベル）成分
広告がなかった時の、
売り上げ水準を表す成分

外生変数（広告フラグ）の係数

状態を2つの成分に分ける
水準成分 と 外生変数の影響

時変係数モデル

今期の状態 ＝ 今期の水準 ＋ 今期の係数 × 広告フラグ

今期の水準 → ランダムウォーク的変動

今期の係数 → ランダムウォーク的変動

水準成分も外生変数の係数も
ともに時間によって変化する

参考：回帰分析との比較

普通の線形回帰モデル

売り上げ ＝ 切片 ＋ 係数 × 広告フラグ

時変係数モデルの状態

今期の状態 ＝ 今期の水準 ＋ 今期の係数 × 広告フラグ

水準 → ランダムウォーク的変動

係数 → ランダムウォーク的変動

切片や、回帰係数がランダムウォーク的に変化することを想定した時変係数モデル

参考：ローカルレベルモデルとの比較

今期の状態 ＝ 前期の水準 ＋ 過程誤差

今期の観測 ＝ 今期の状態 ＋ 観測誤差

ローカルレベルモデルは、
水準成分のみを持つモデルであるといえる

用語：水準と係数／確定的と確率的

確定的水準

通常の線形回帰モデルの「切片」のように、
時間によって変化しない水準成分

確率的水準

ローカルレベルモデルのように、
時間によって確率的に変化する水準

確定的係数

通常の線形回帰モデルの「回帰係数」

確率的係数（時変係数）

時間によって確率的に変化する係数

時変係数モデル

今期の水準 ＝ 前期の水準 ＋ 過程誤差①

今期の係数 ＝ 前期の係数 ＋ 過程誤差②

0：広告なし／1：広告あり

今期の観測 ＝ 今期の水準 ＋ 今期の係数 × 広告フラグ

＋ 観測誤差

時変係数モデル

ex_t:時点tでの外生変数
β_t：時点tでの回帰係数

状態方程式

$$\mu_t = \mu_{t-1} + w_t, \qquad w_t \sim \mathrm{N}(0, \sigma_w^2)$$

$$\beta_t = \beta_{t-1} + \tau_t, \qquad \tau_t \sim \mathrm{N}(0, \sigma_\tau^2)$$

$$\alpha_t = \mu_t + \beta_t \cdot ex_t$$

観測方程式

$$y_t = \alpha_t + v_t, \quad v_t \sim \mathrm{N}(0, \sigma_v^2)$$

内容

1. 時変係数モデルの理論
2. 時変係数モデルの推定

Rを用いた実演