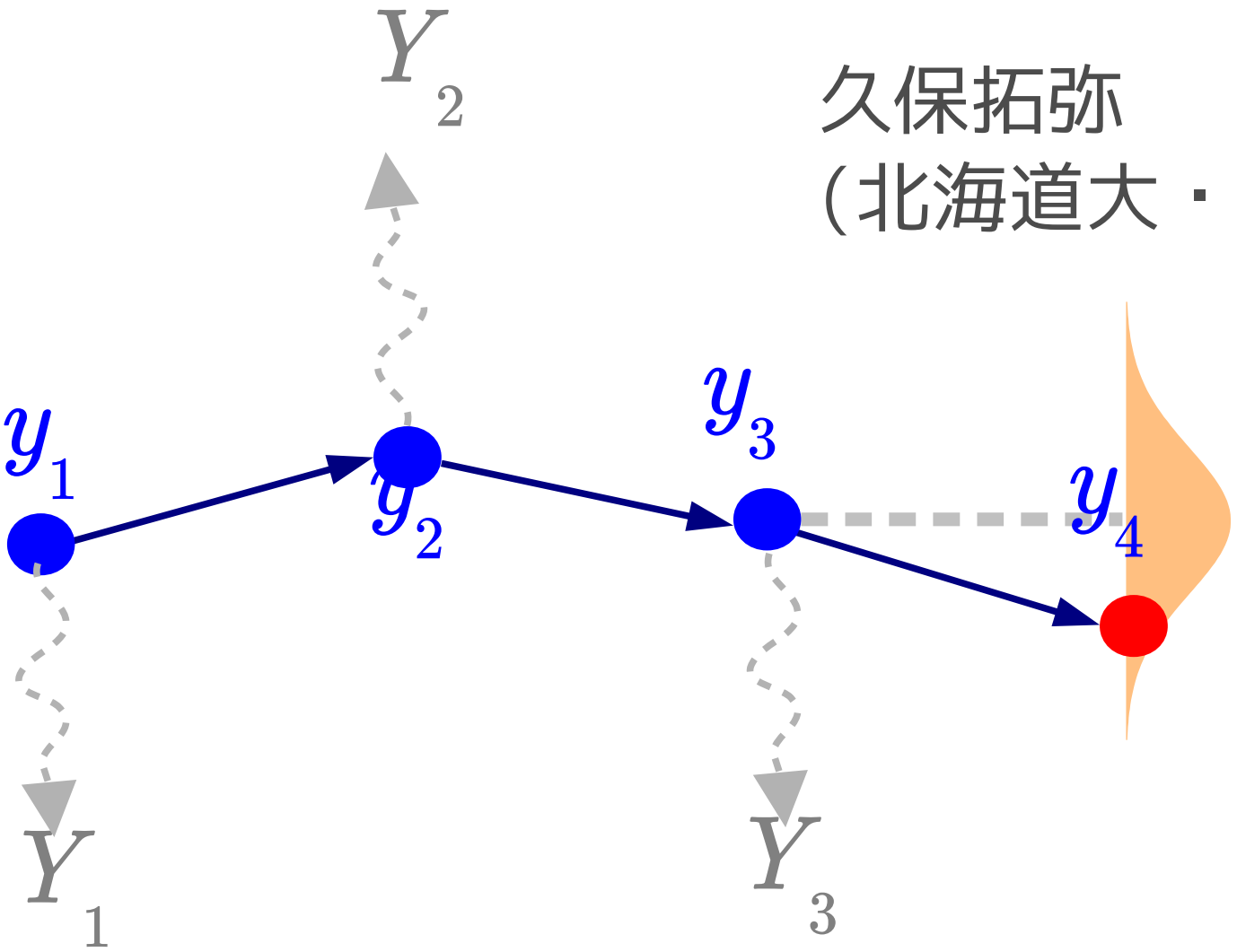
時系列データ解析のための状態空間モデル（続）

久保拓弥
（北海道大・環境科学）



生態学会・仙台大会の時系列なモデル

他にもあったらすみません…

- [T01] 階層的なプロセスをモデル化する：階層モデルによる生態データ解析
- [T09] 不確実性下の哺乳類管理：管理施策の選択とその課題
- [W25]（この集会）粕谷さん「うたがわしい回帰」
- [T23] 生態学における因果推論：convergent cross mapping とその周辺
- [P2-289] イモゾウムシ個体数密度の寄主植物による違い–ベイズ統計モデリングによる解析（本間さん・高倉さん）

今回，説明してみたいこと

- 時系列データ:単純な回帰はダメ(続)
- 状態空間モデル: 乱歩と雑音の分離
- 差分と時間的自己相関係数
- 欠測と不等間隔
- 時系列と「対応のある」データ
- 説明しないこと - 因果推定など

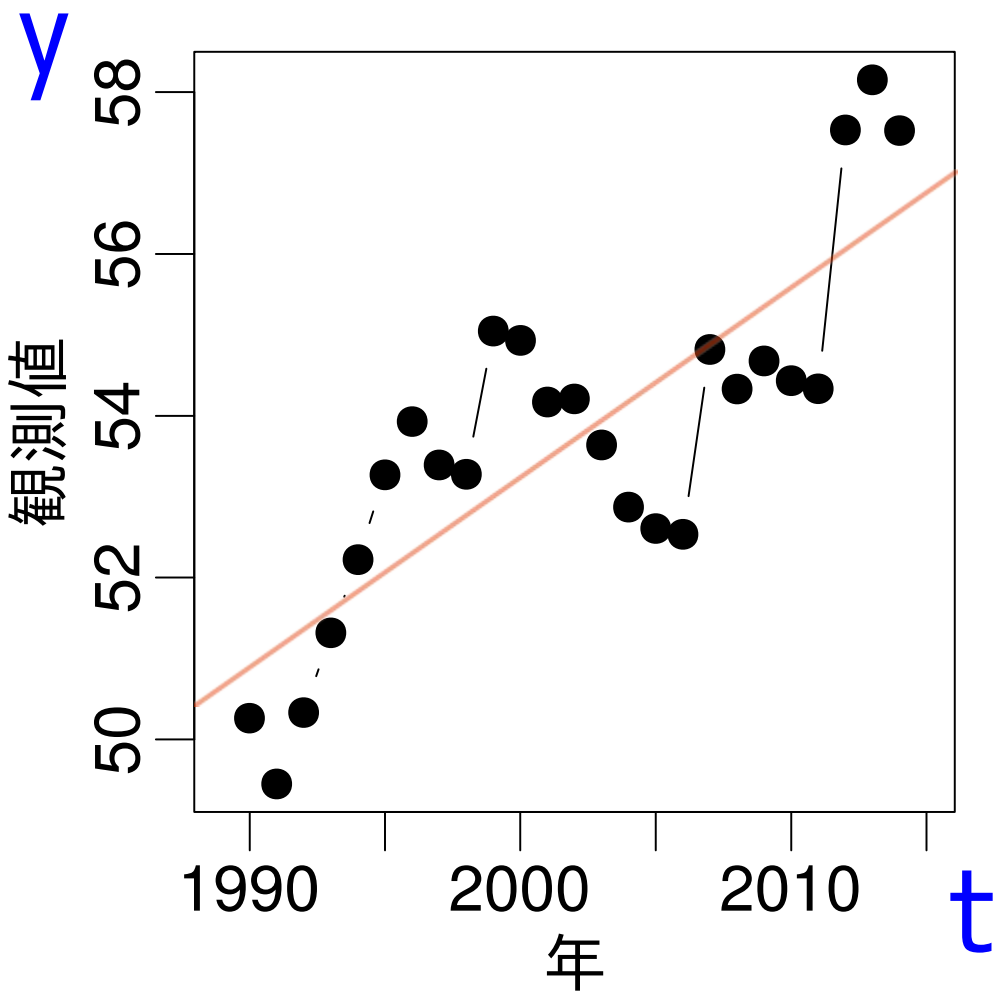
今日の要点

時系列データの解析は

階層ベイズモデル化した

状態空間モデルを使うのが便利

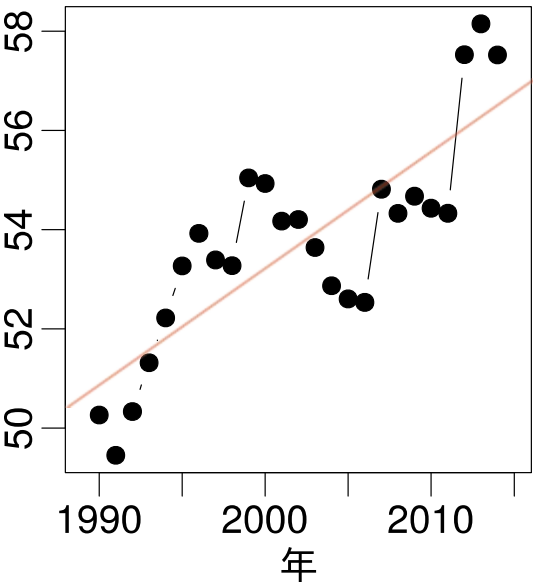
時間相関のある時系列データに…



glm(y ~ t)

…と，モデルを
あてはめてみた

「やったーゆーいだ!!」……??

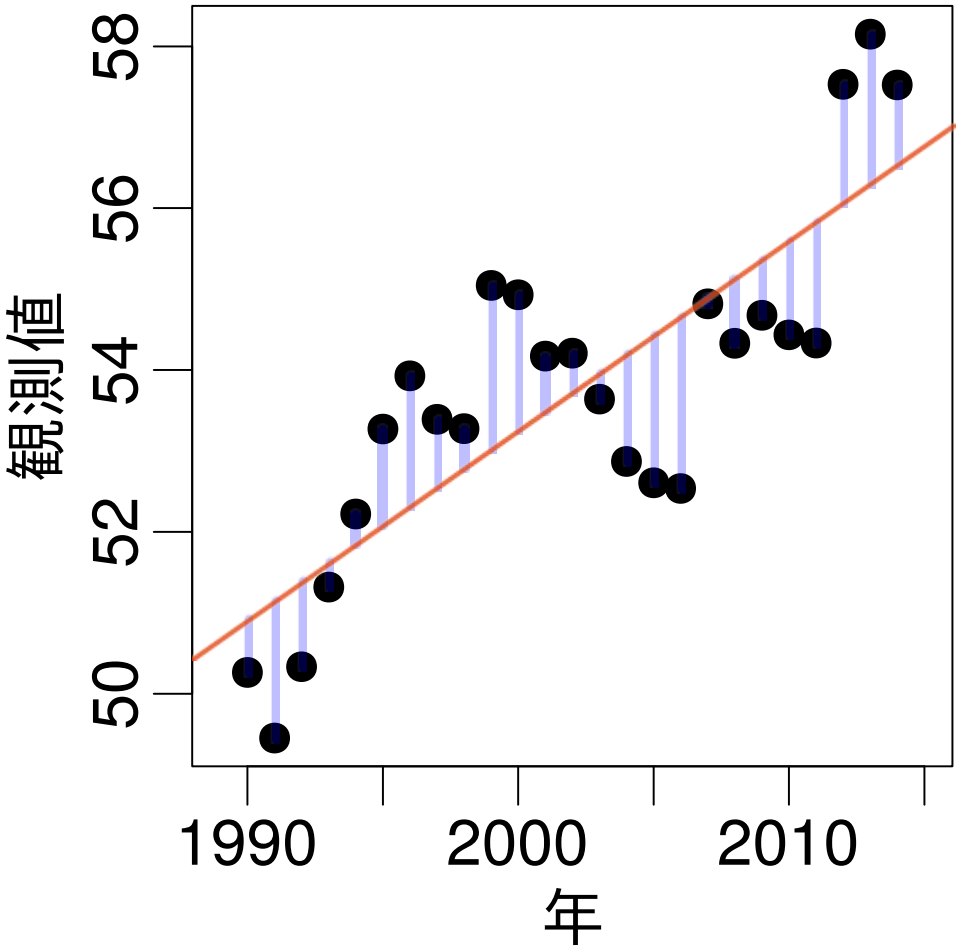



```
> summary(glm(formula = y ~ t))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1295  -1.0583  -0.0817   0.9860   2.0188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -414.5655    71.4761   -5.80  6.6e-06
t              0.2339     0.0357    6.55  1.1e-06
```

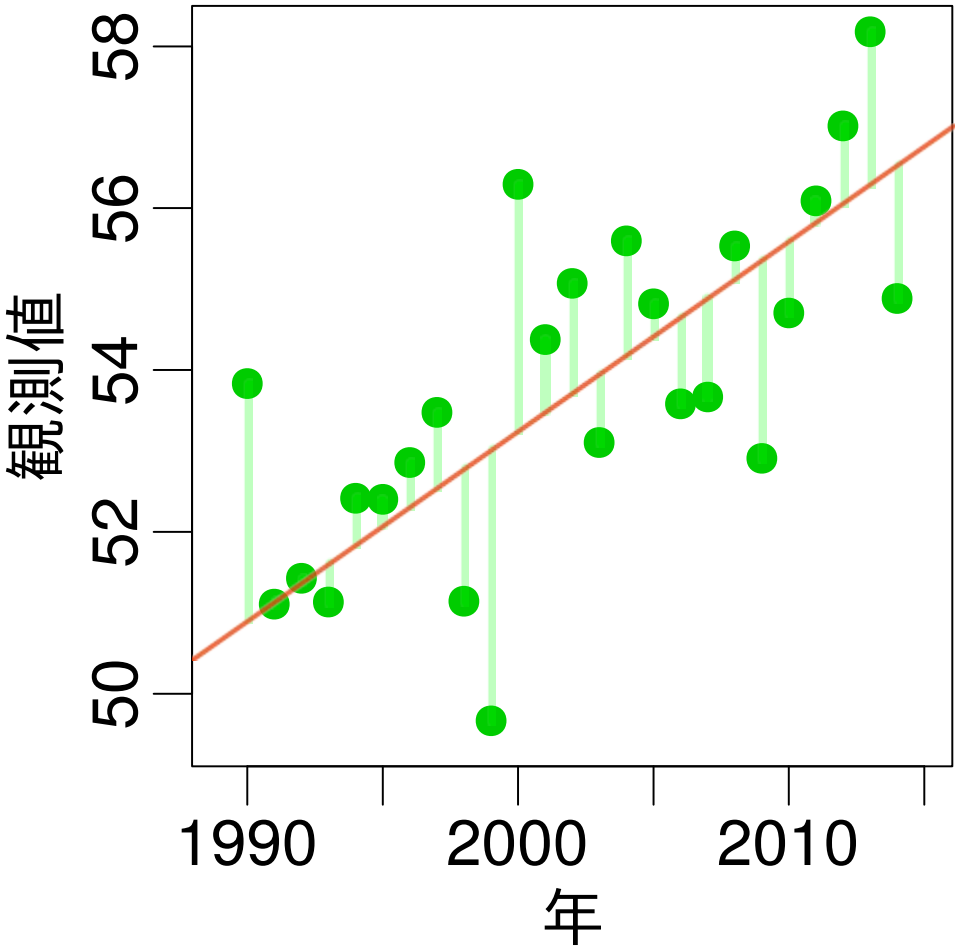
これはまちがい→ glm(時系列Y ~ 時間 t)

統計モデルがおかしい?

時系列の「ずれ」



GLM のずれ



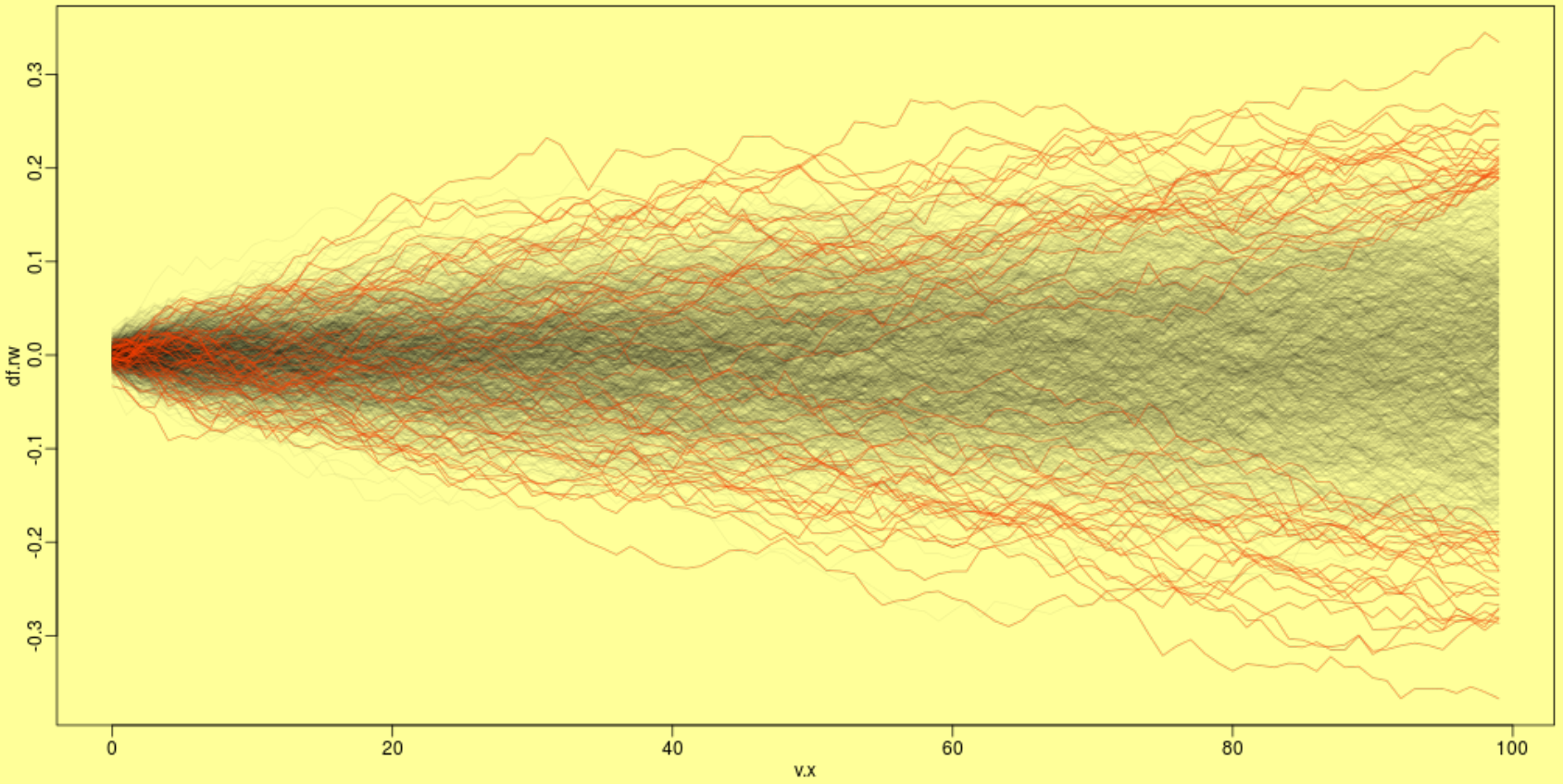
直線からのずれがちがう!

時間的自己相関がある

時間的自己相関がない

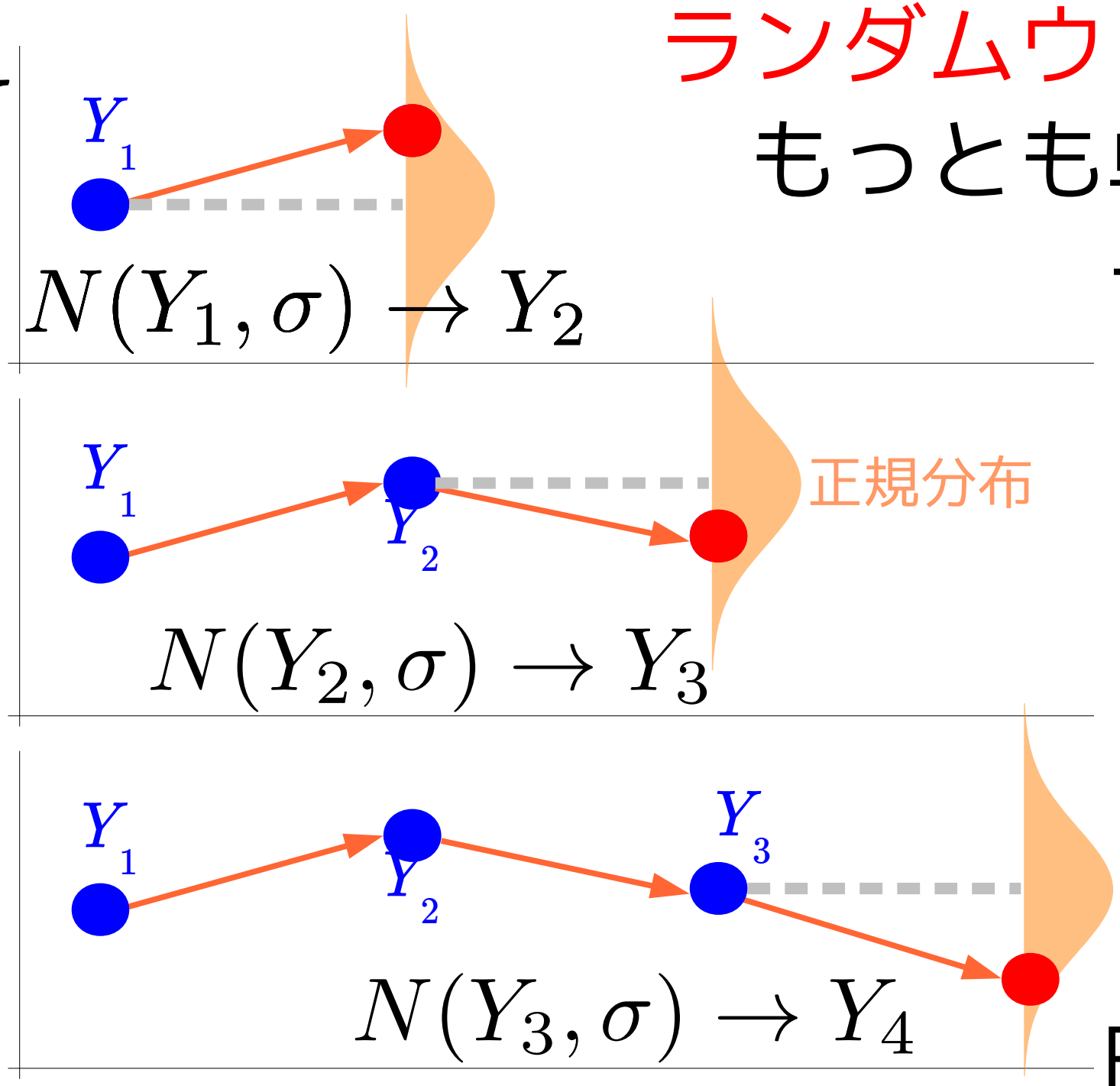
時系列の基本モデルのひとつ

ランダムウォーク（乱歩）

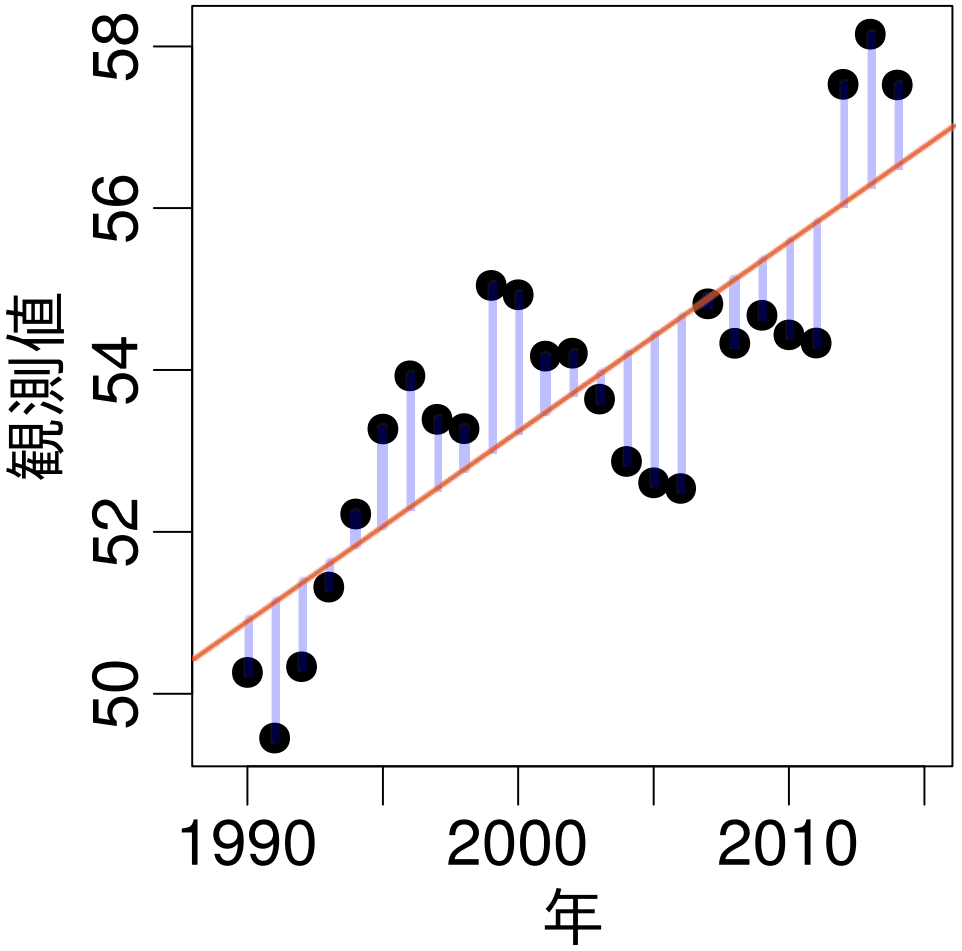


ランダムウォーク

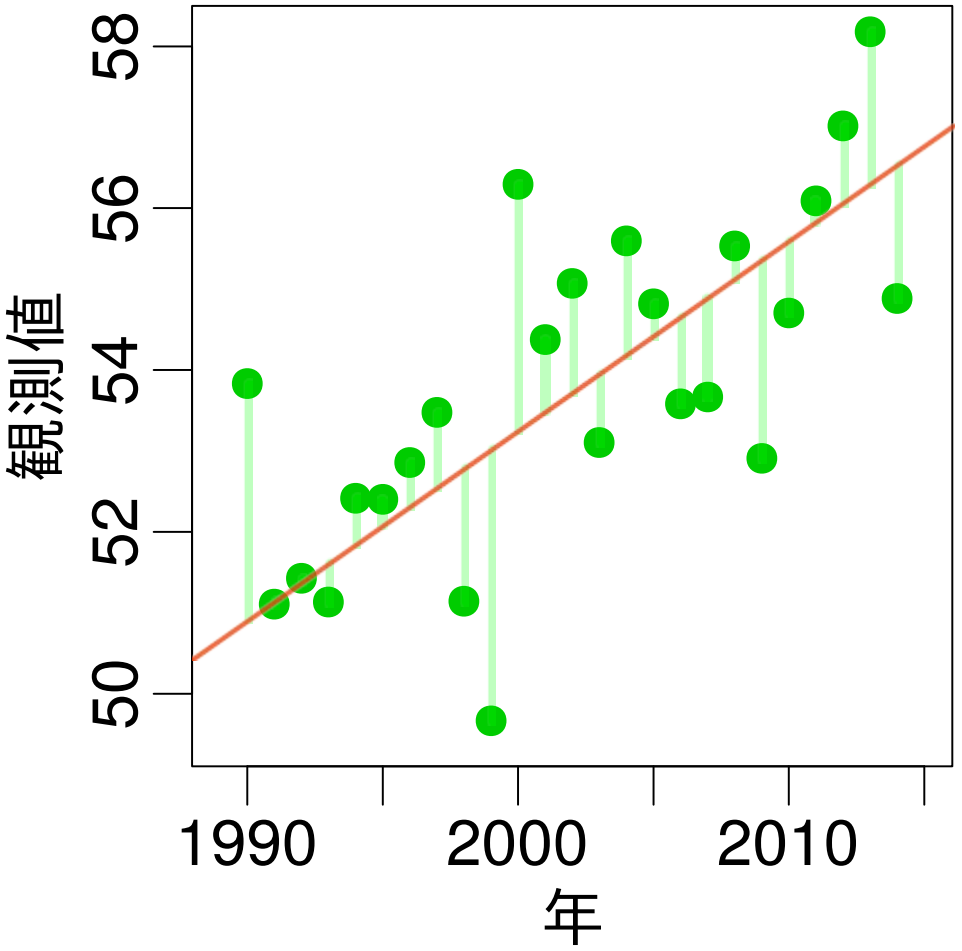
もっとも単純なモデル



時系列の「ずれ」



GLM のずれ



直線からのずれがちがう!

時間的自己相関がある

時間的自己相関がない

時間的自己相関

（略称:自己相関，時間相関）

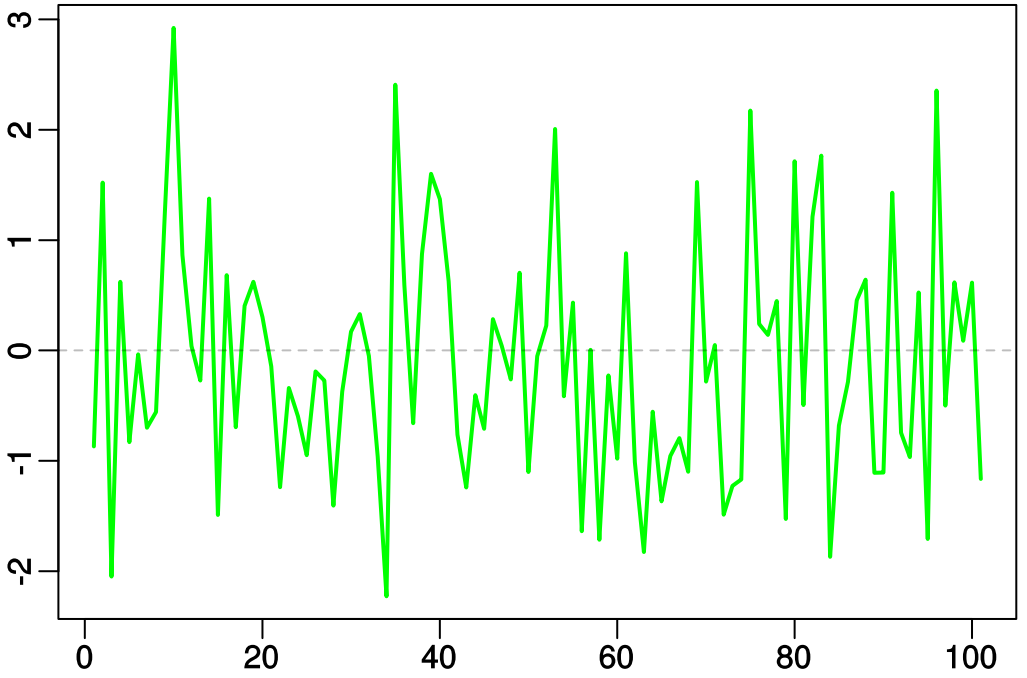
を調べたらいいの?

$$\rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-k})}}$$

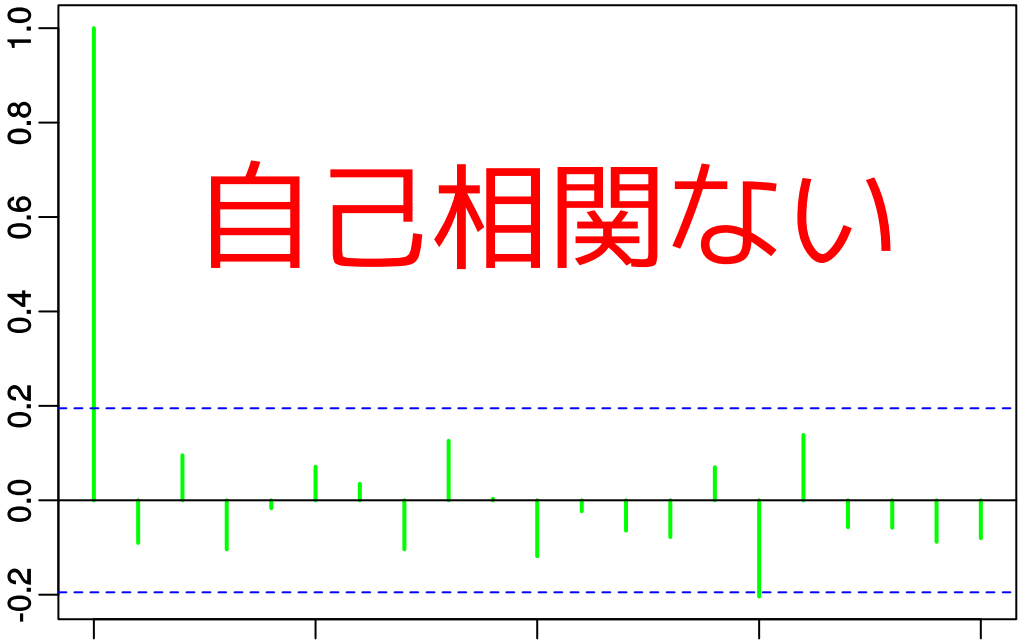
R の ts クラス: 時系列をあつかう

```
plot(ts(Y))
```

これはたんなる
100 個の正規乱数



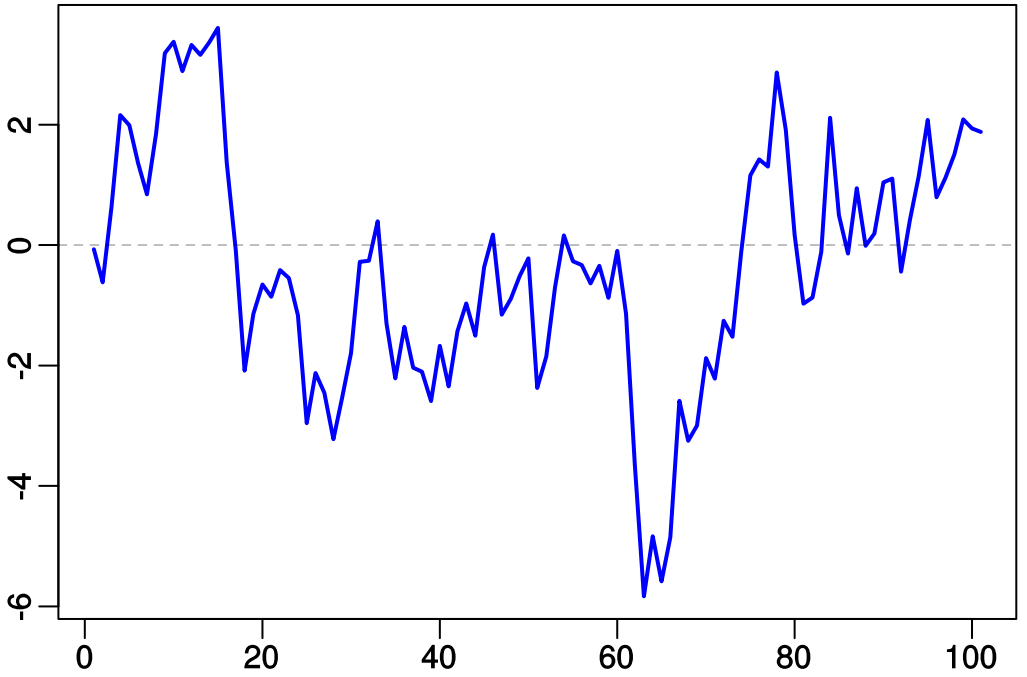

```
plot(acf(ts(Y)))
```



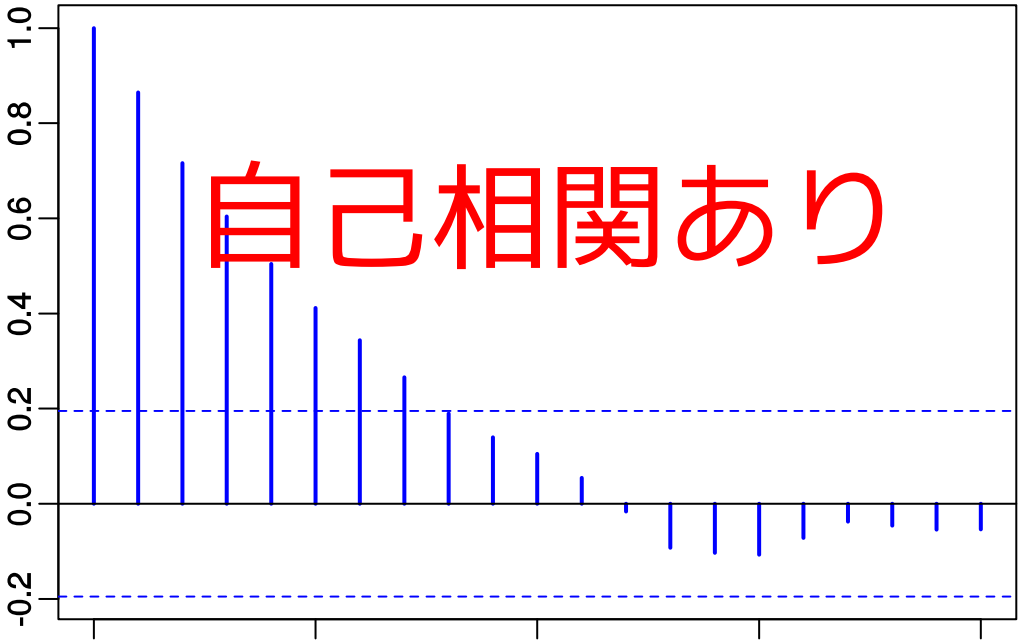
自己相関ない

自己相関減衰の様子を図示

```
plot(ts(Y))
```




```
plot(acf(ts(Y)))
```



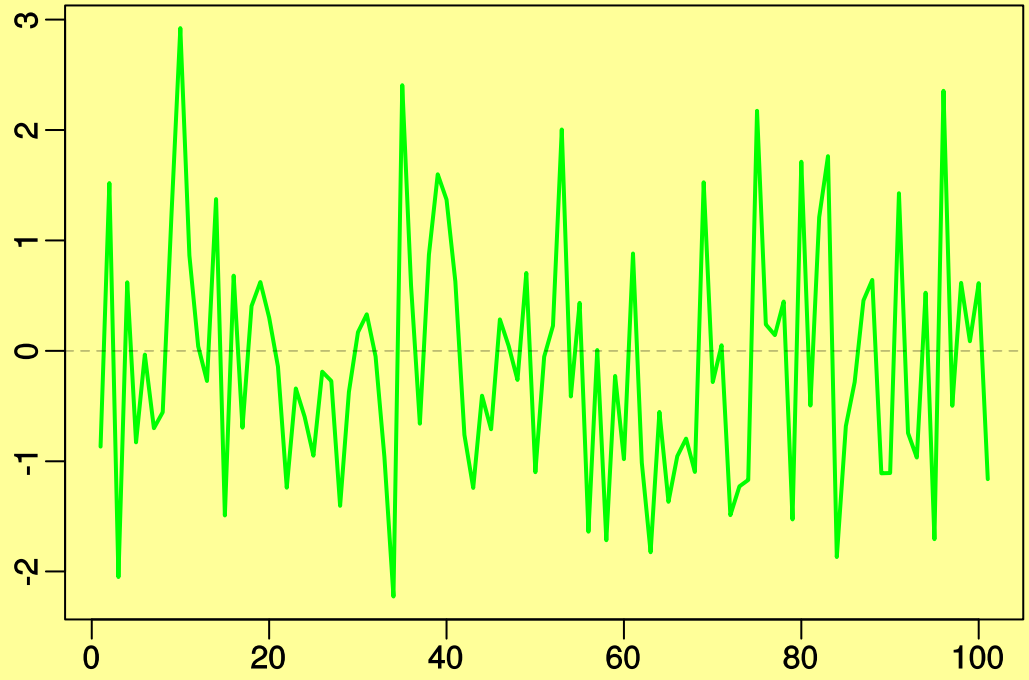
自己相関あり

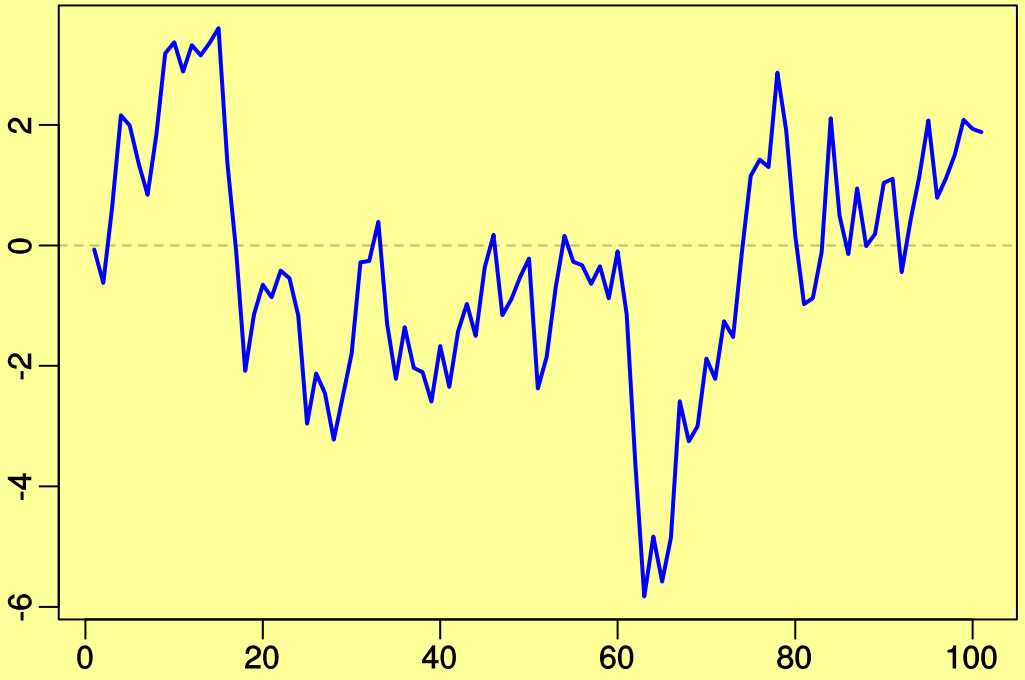
状態空間モデルでたちむかう

時系列データ解析

いろいろな時系列データを

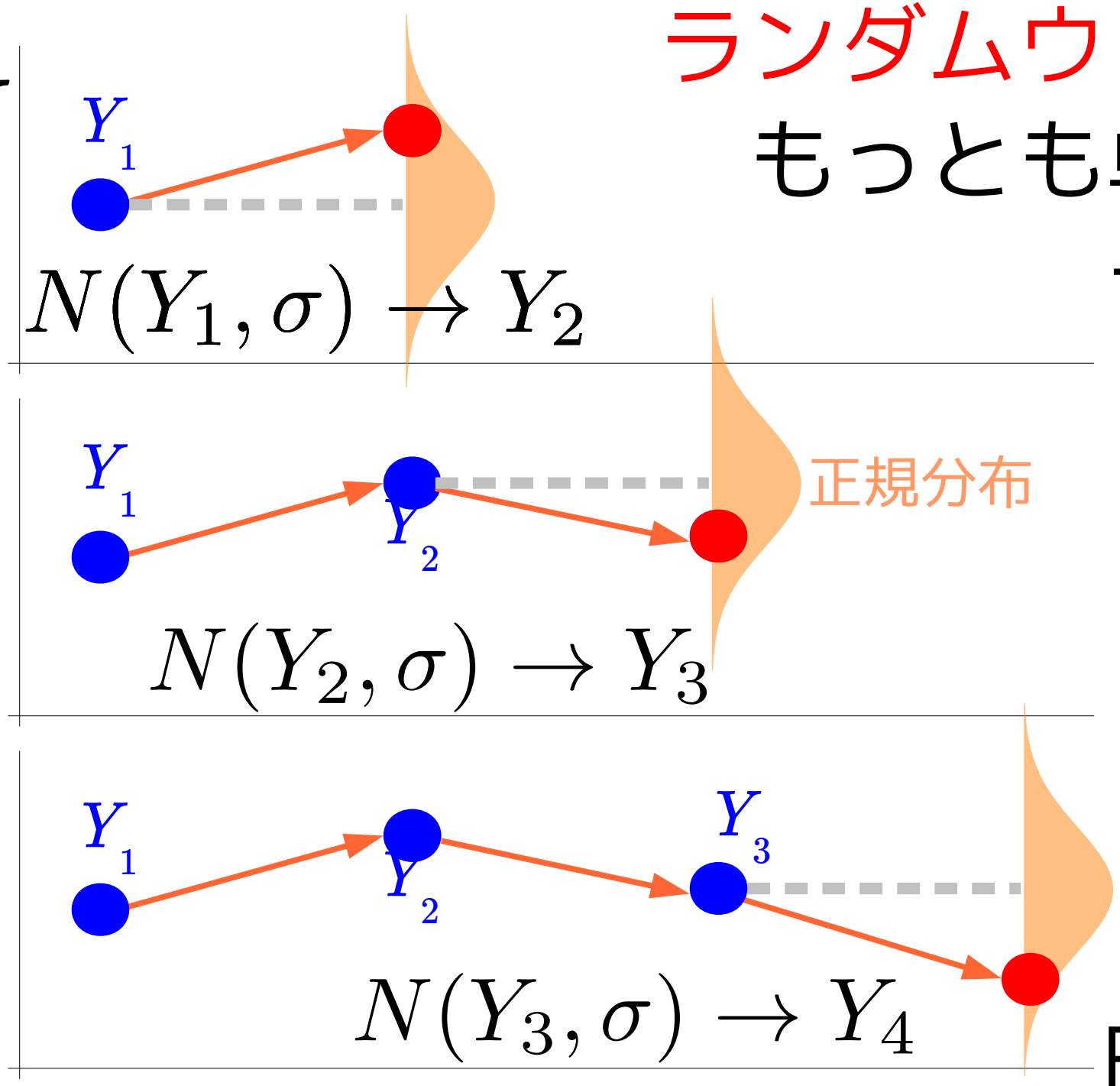
統一的にあつかえないか?





ランダムウォーク

もっとも単純なモデル



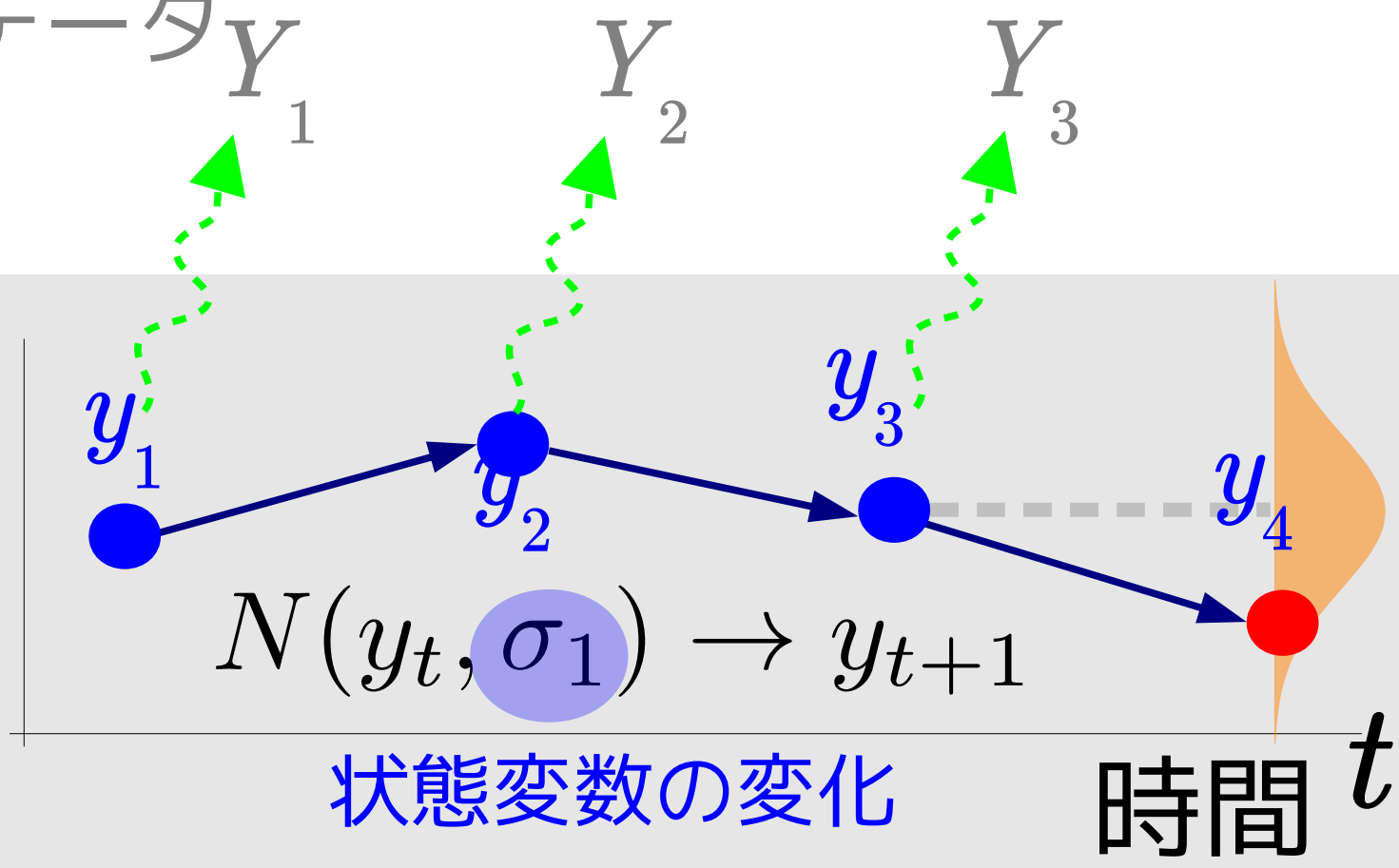
状態空間モデル

二種類のσをもつ

観測の誤差

$$N(y_t, \sigma_2) \rightarrow Y_t$$

観測データ



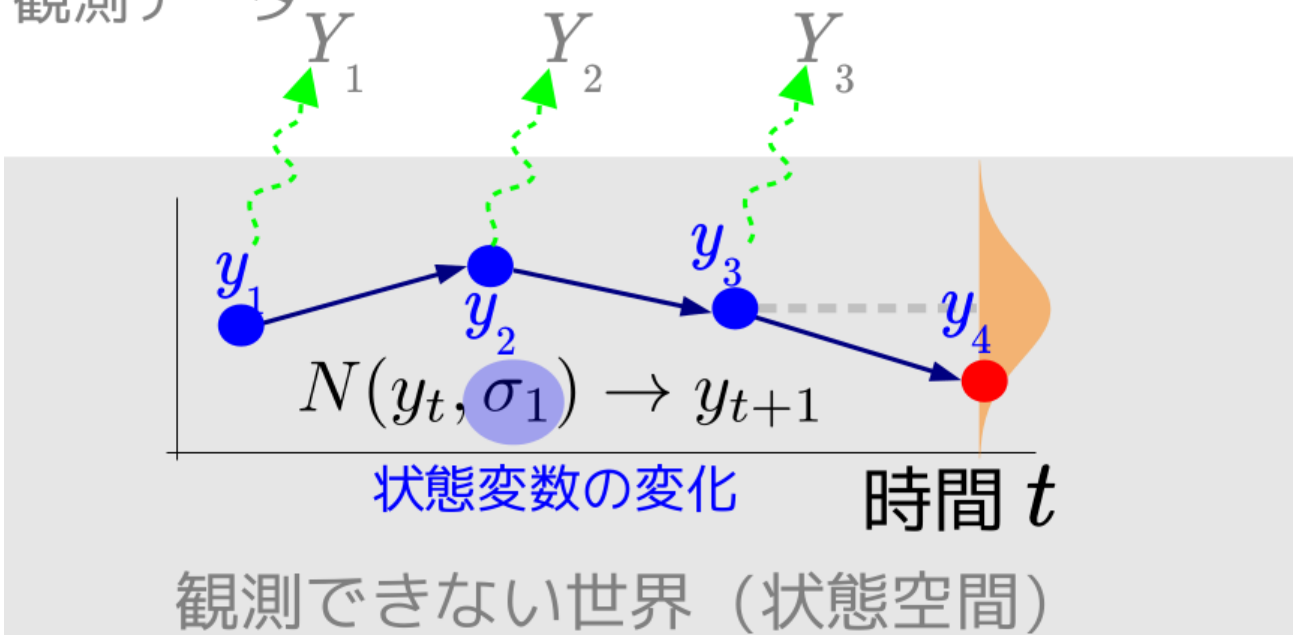
観測できない世界（状態空間）

状態空間モデル

観測の誤差

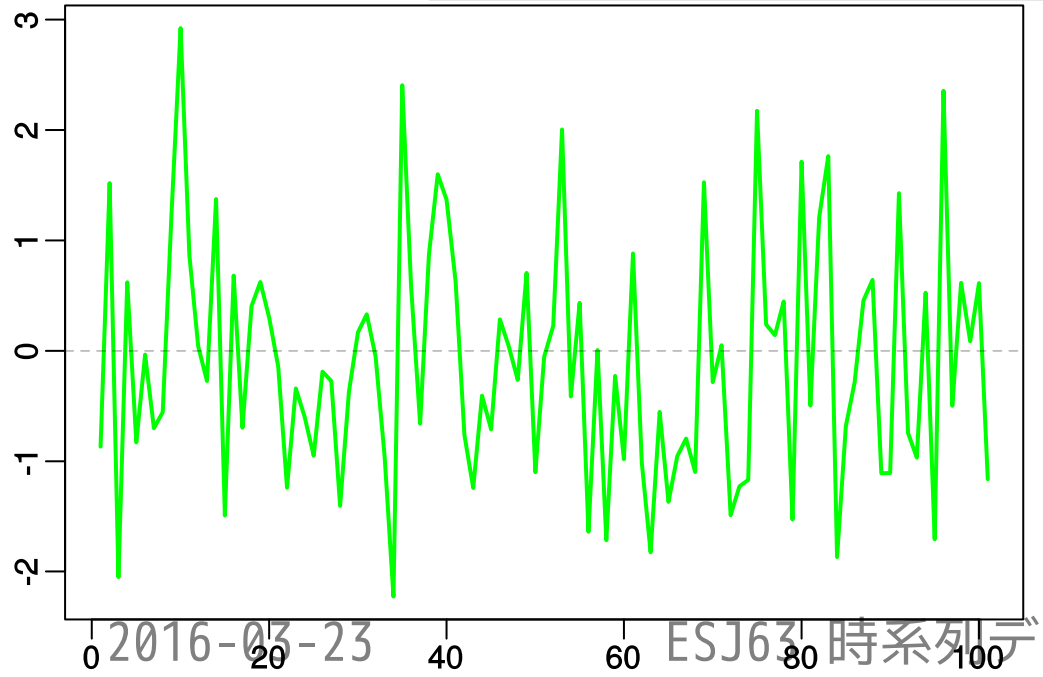
$N(y_t, \sigma_2) \to Y_t$ 二種類の σ をもつ

観測データ

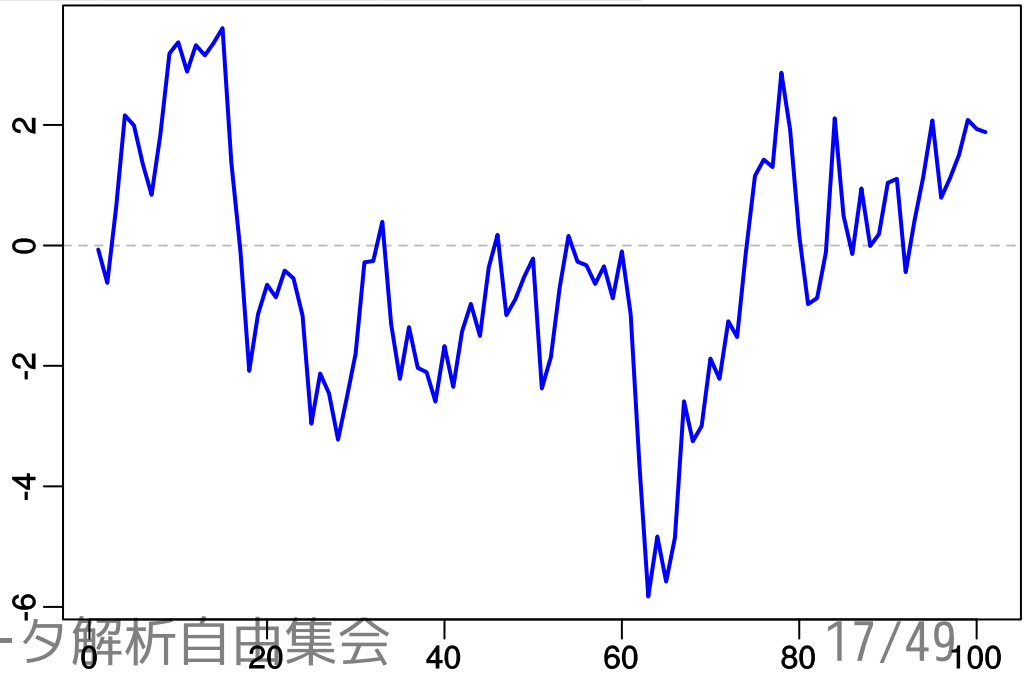


観測できない世界（状態空間）

σ_2 大
σ_1 小



σ_2 小
σ_1 大



状態空間モデル + 観測モデル

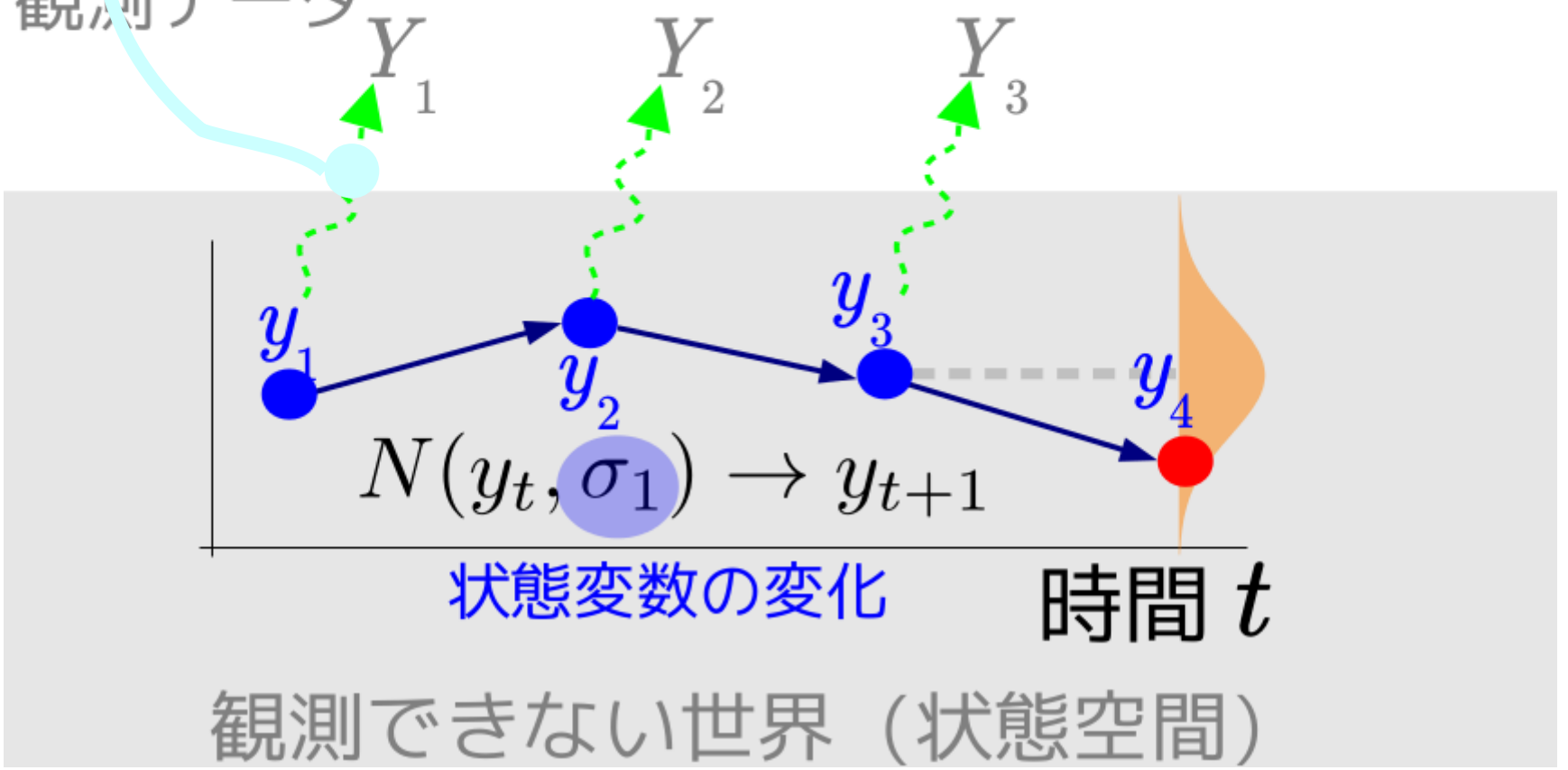
この部分にポアソン分布や
二項分布をいれる

誤差

状態空間モデル

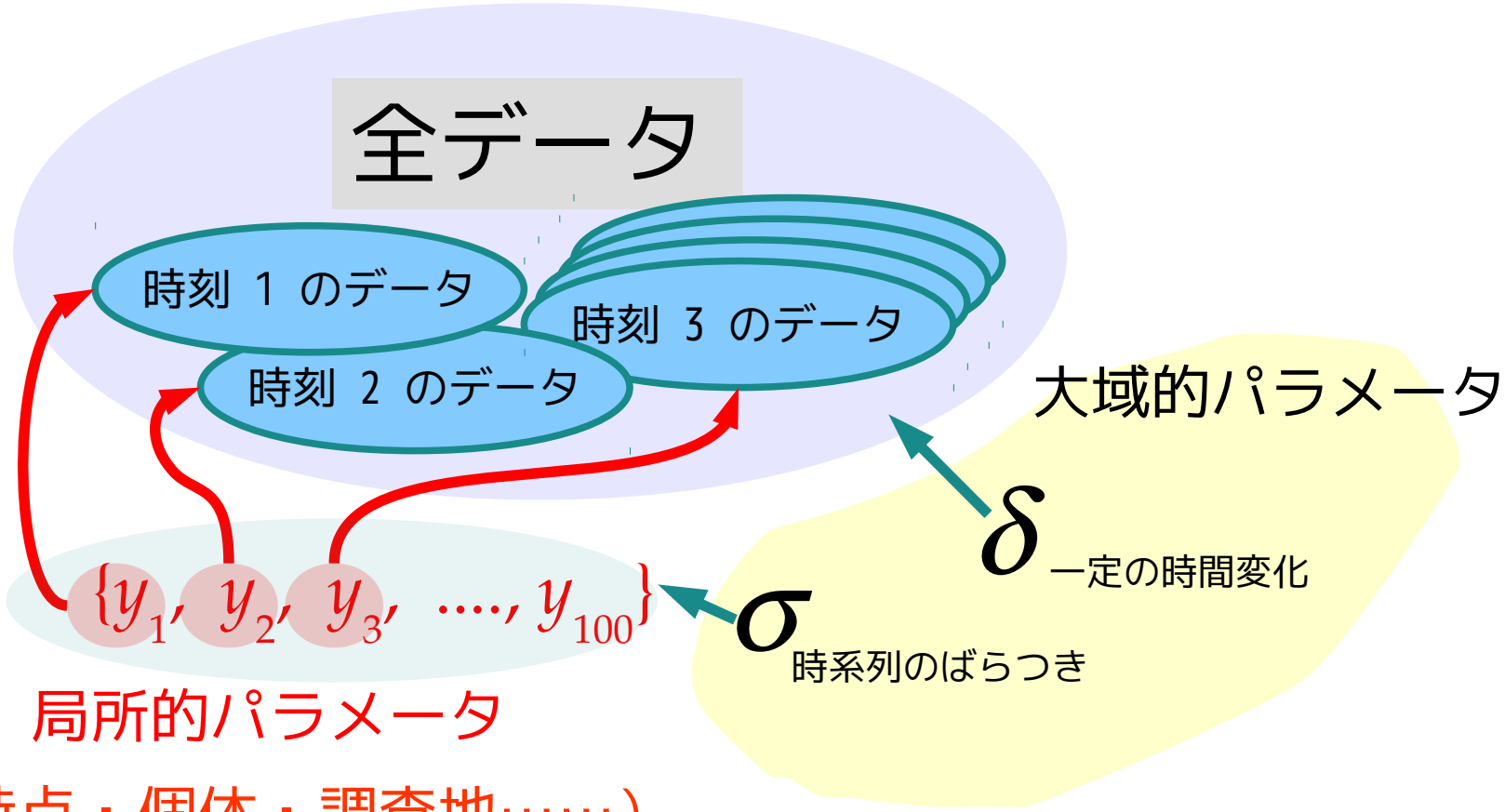
$N(y_t, \sigma_2) \rightarrow Y_t$ 二種類の σ をもつ

観測データ



状態空間モデルは階層ベイズモデル

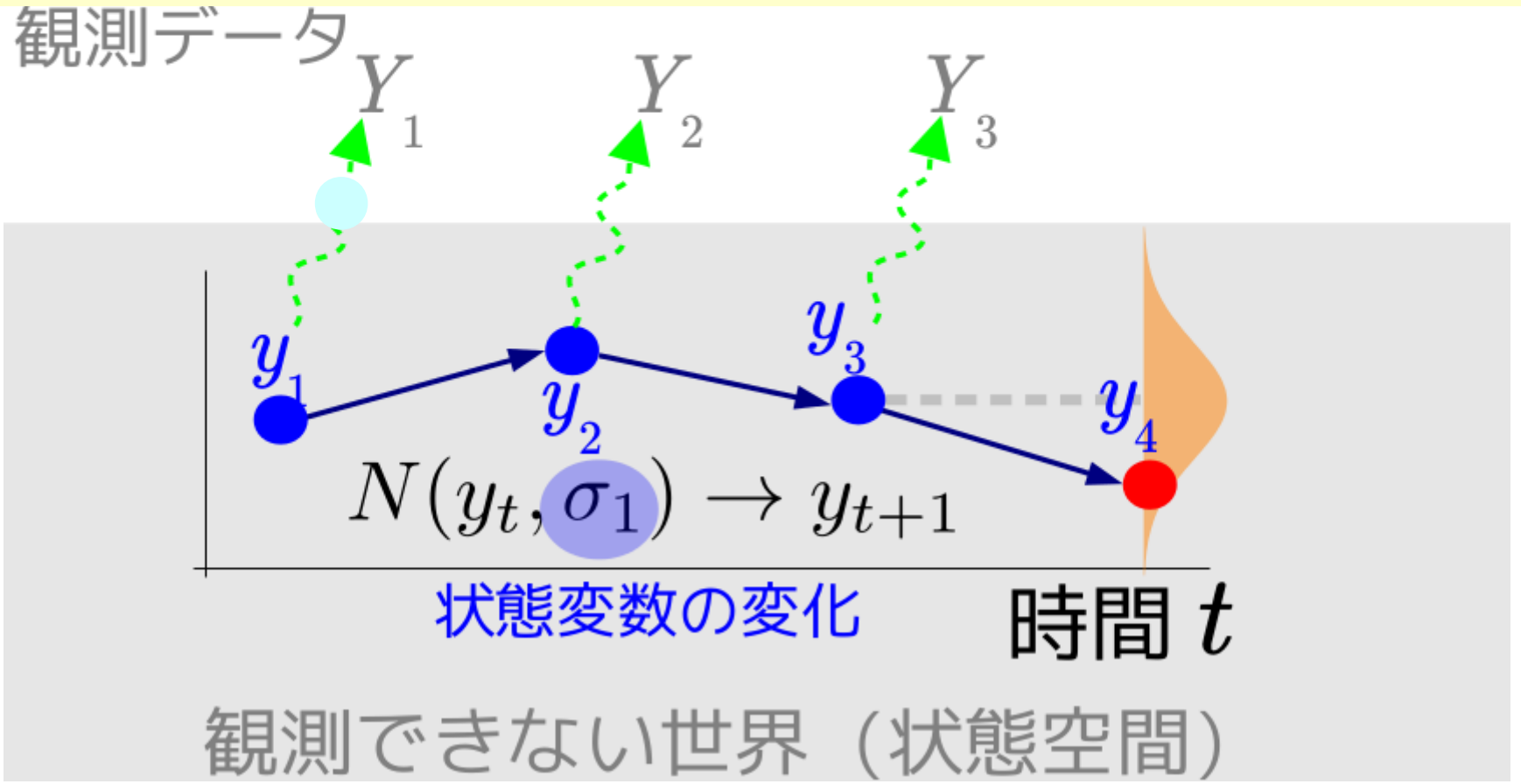
多数の「似たようなパラメーター」たちに
「適切」な制約を加えて推定できる



（たくさんの時点・個体・調査地……）

状態空間モデル + 観測モデル

他にも季節変動などを
入れることができます



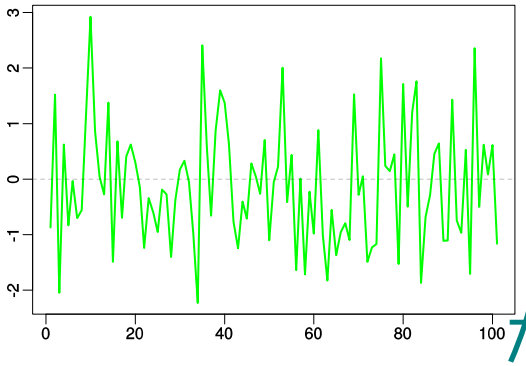
時間的自己相関係数

はいつも役にたつわけではない?

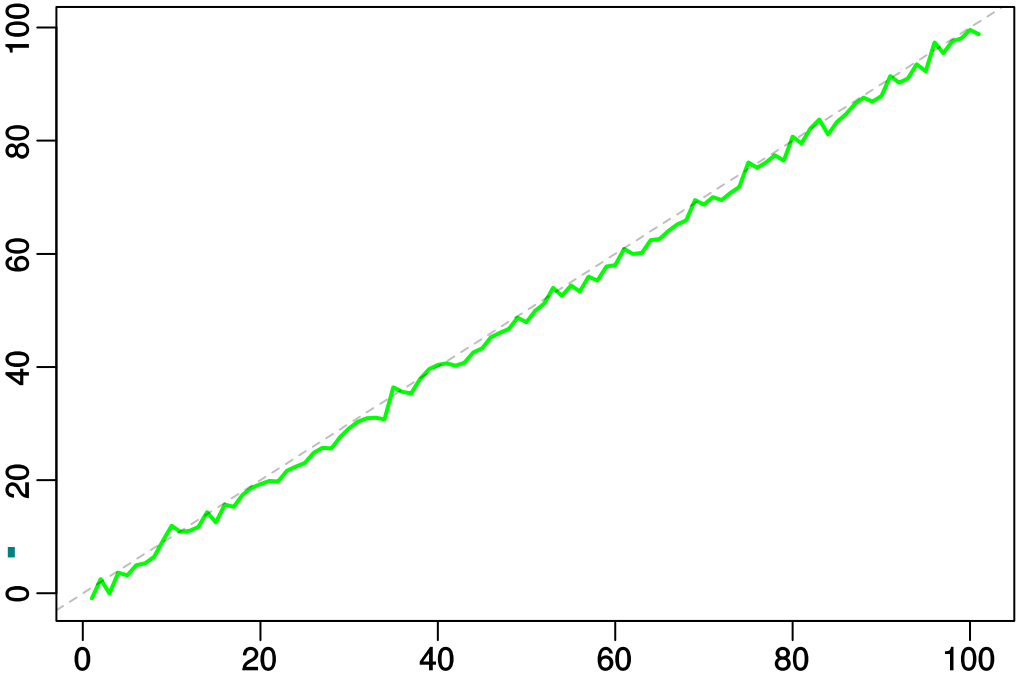
$$\rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-k})}}$$

各点独立のデータをナナメにすると?

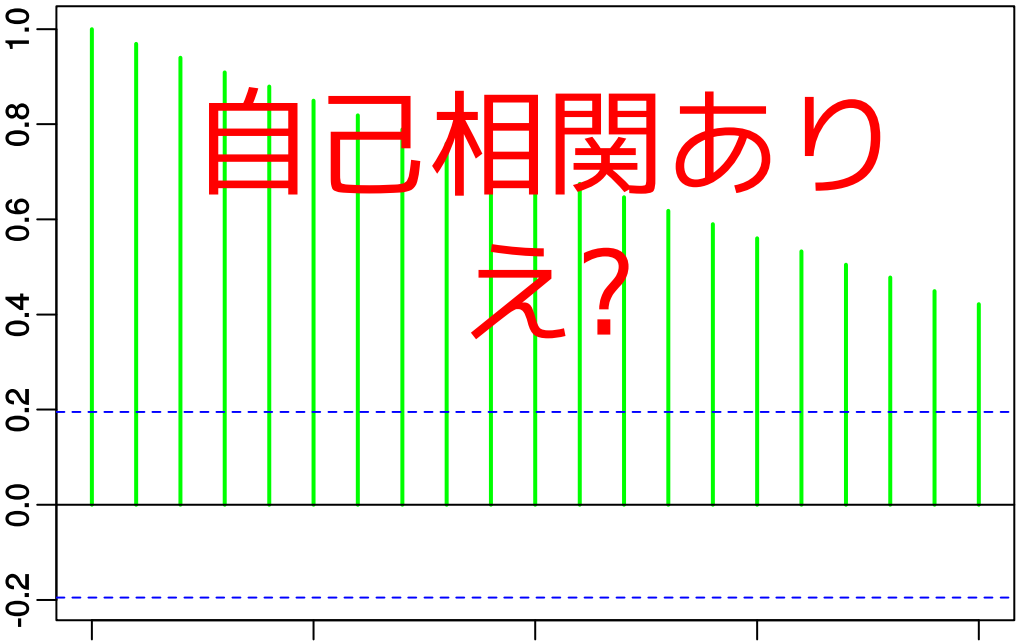
```
plot(ts(Y))
```



これを
ナナメに
したもの
なんだけど…



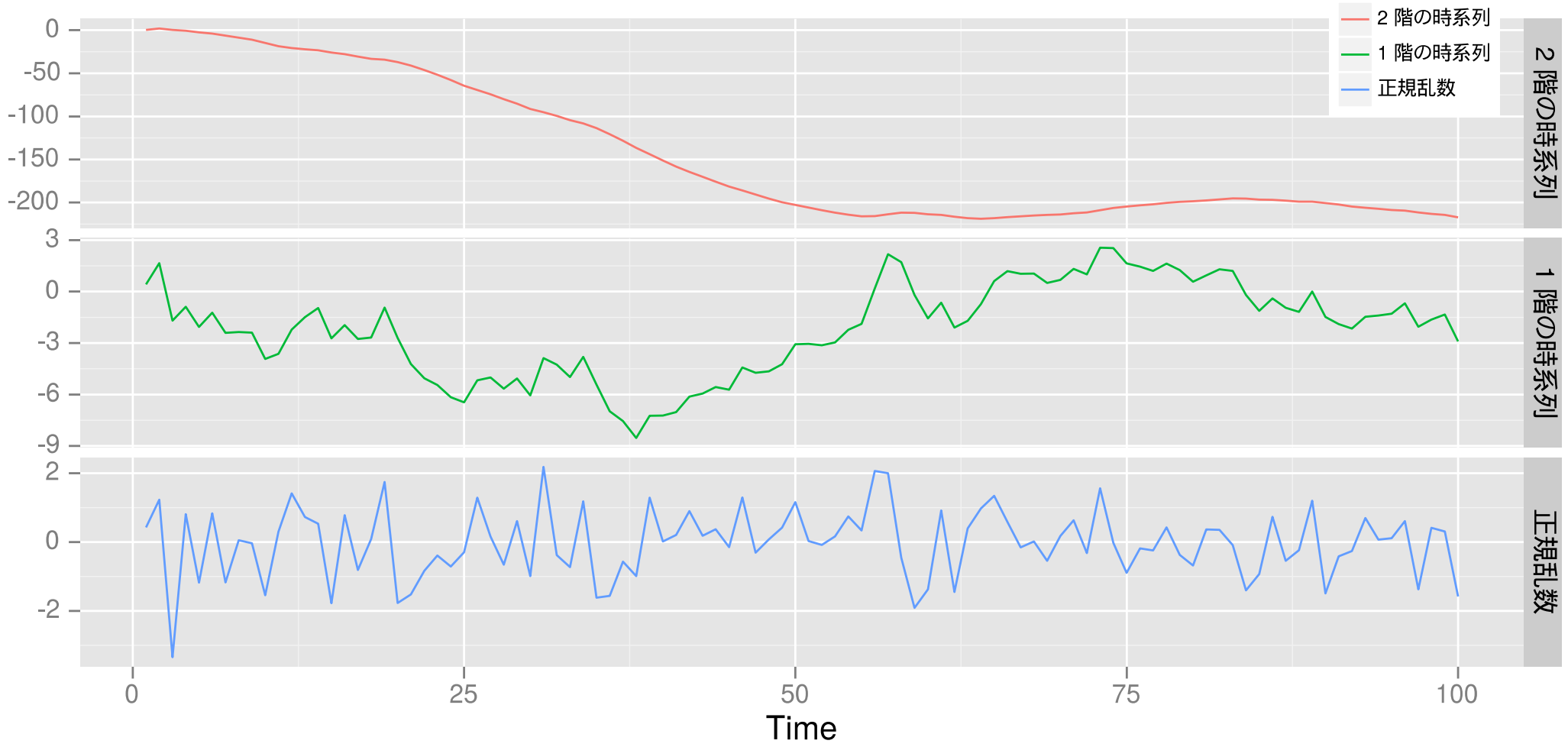

```
plot(acf(ts(Y)))
```



自己相関あり
え?

時系列データの「差分」をみよう

自己相関係数もいいけど差分を調べるのが基本



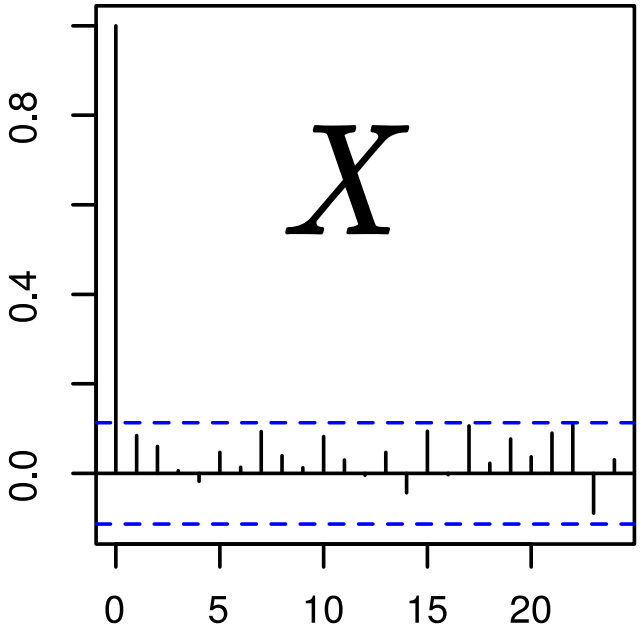
各時刻で独立なノイズで隠される

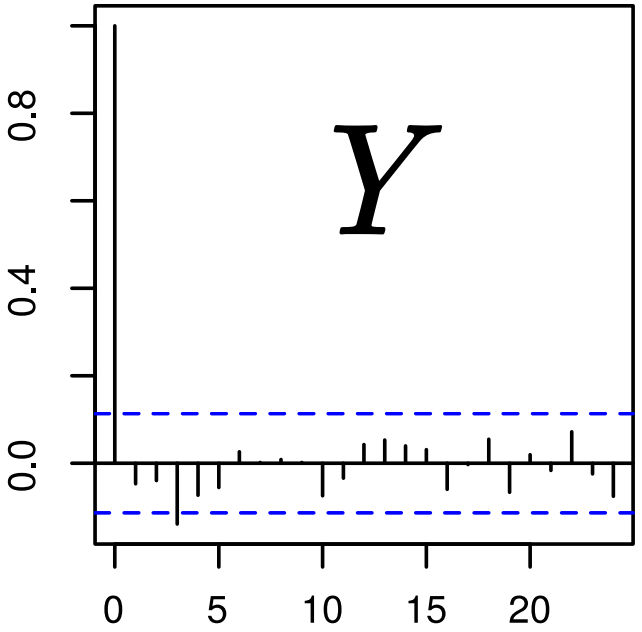
時間的自己相関係数

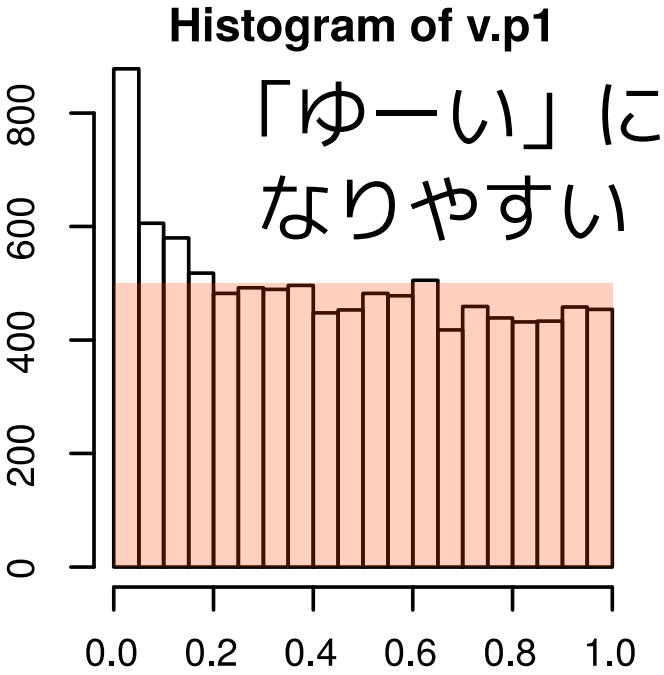
$$\rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-k})}}$$

ノイズの大きな時系列にうもれたワナ？

時間的自己相関のない時系列？







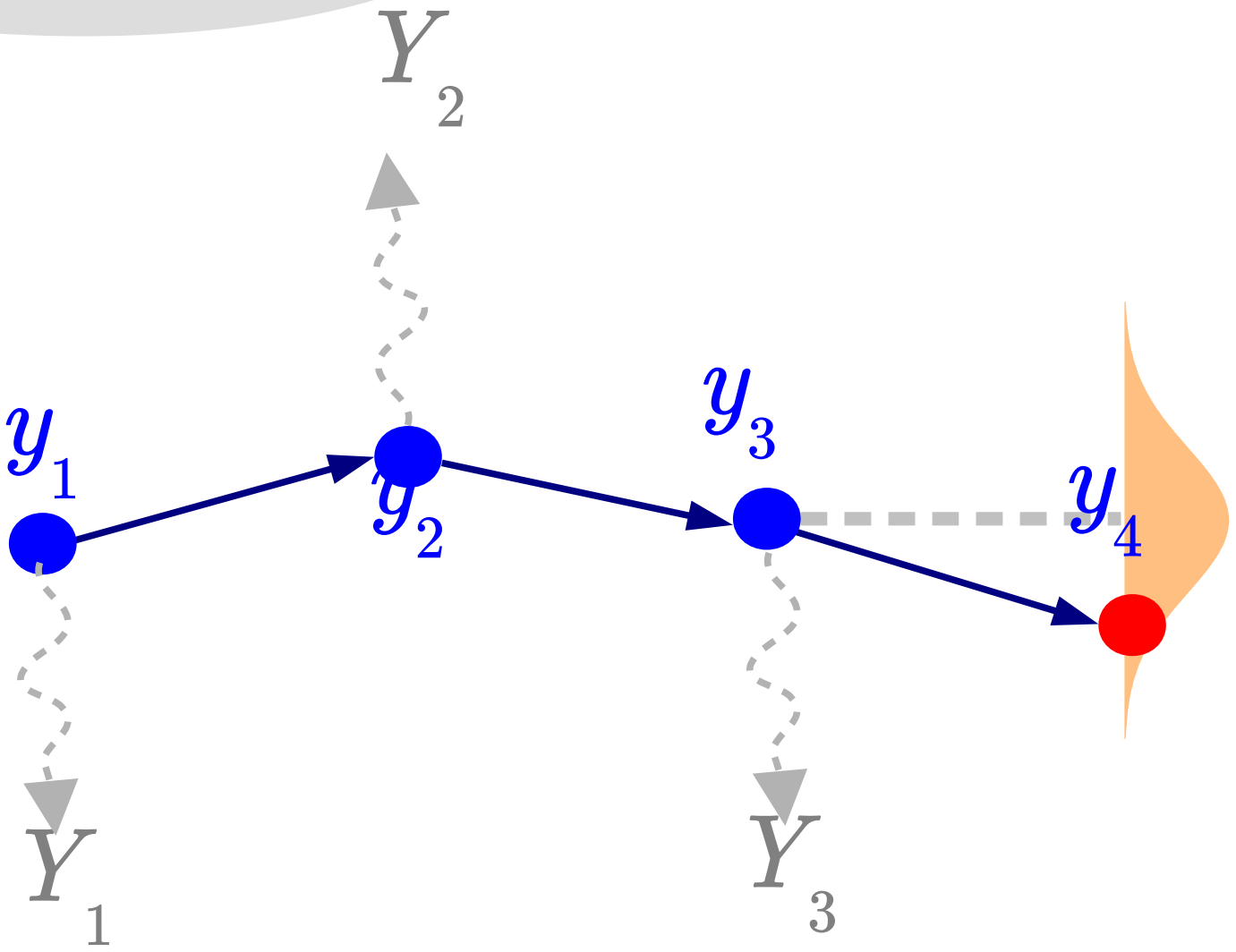
しかし glm(Y ~ X) とすると…

状態空間モデル

二種類のσをもつ

観測の誤差 あるいは

各時刻で独立なノイズ



状態空間モデルを R で使うには

どうすればよいか？

``関数派'' vs ``BUGS 派''？

どうやてモデルをあてはめる?

R の状態空間モデルの

package いろいろある

```
library(dlm)
```

```
library(KFAS)
```

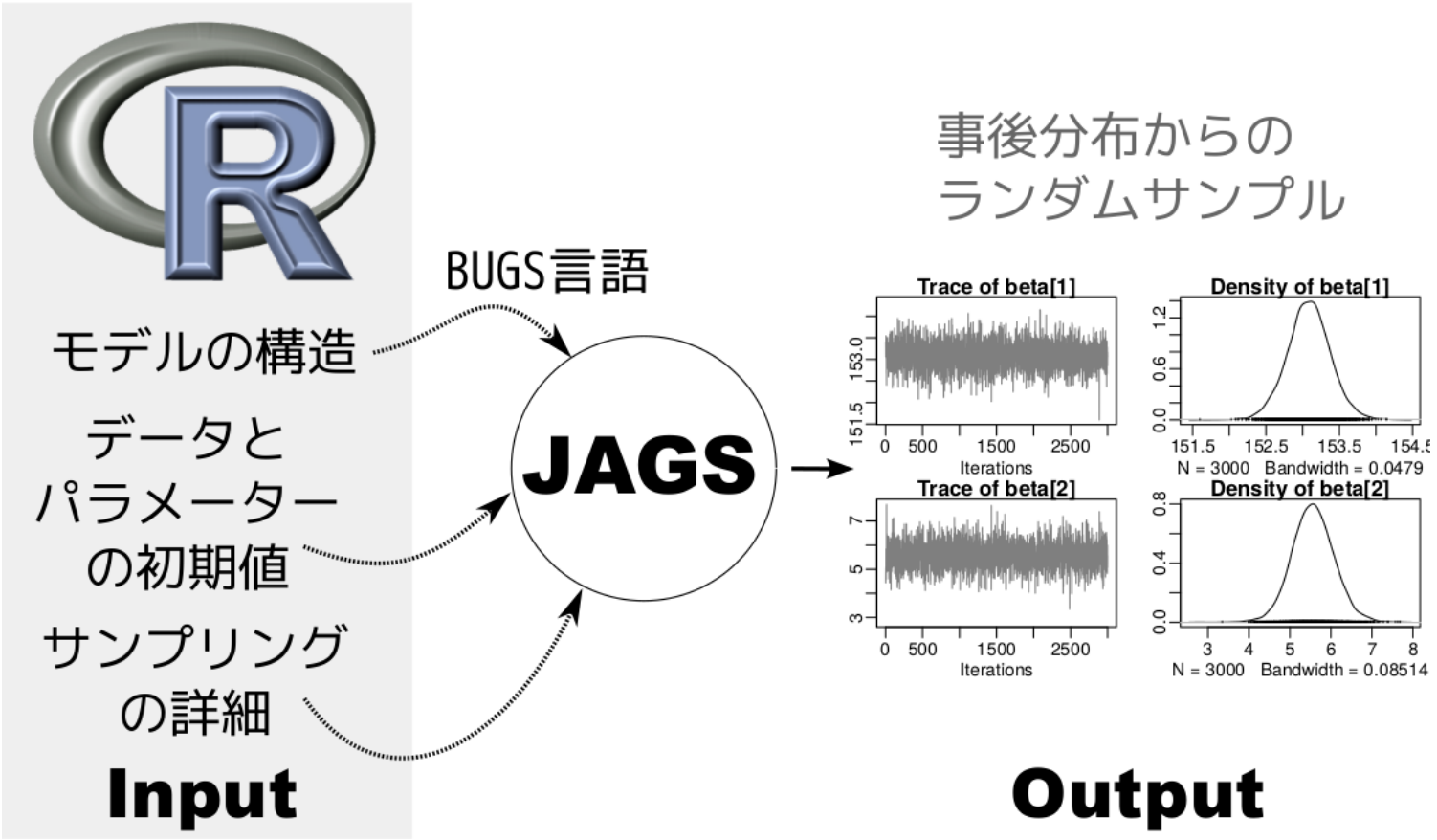
しかしより一般化したモデルに

ついての理解が必要かも

たとえば JAGS で

BUGS 言語でこの単純な

階層ベイズモデルを記述できる



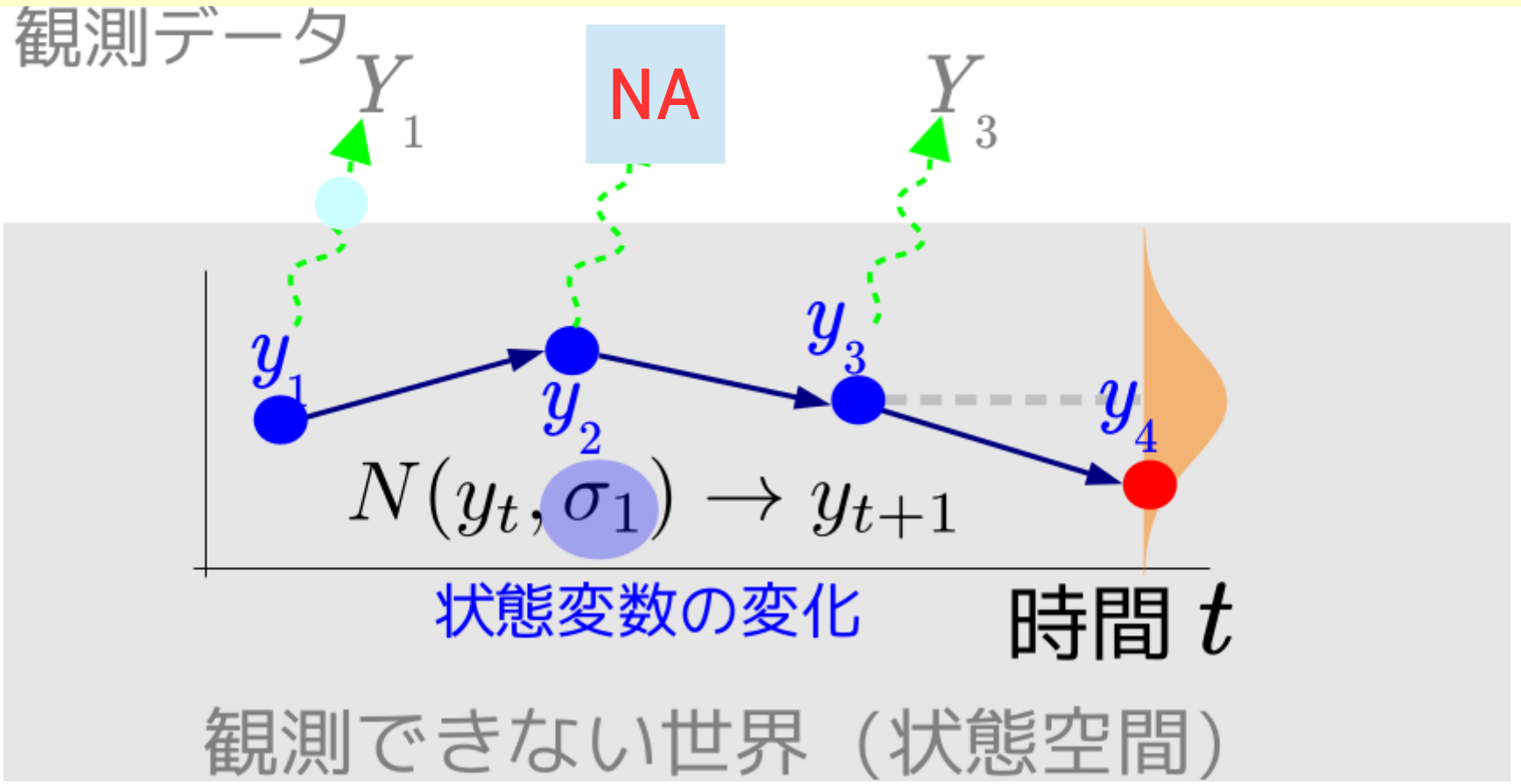
```
model
{
    Tau.Noninformative <- 0.0001
    Y[1] ~ dnorm(y[1], tau[2])
    y[1] ~ dnorm(0, Tau.Noninformative)
    for (t in 2:N.Y) {
        Y[t] ~ dnorm(y[t], tau[2])
        y[t] ~ dnorm(m[t], tau[1])
        m[t] <- delta + y[t - 1]
    }
    delta ~ dnorm(0, Tau.Noninformative)
    for (k in 1:2) {
        tau[k] <- 1 / (s[k] * s[k])
        s[k] ~ dunif(0, 10000)
    }
}
```

状態空間モデルを使う利点

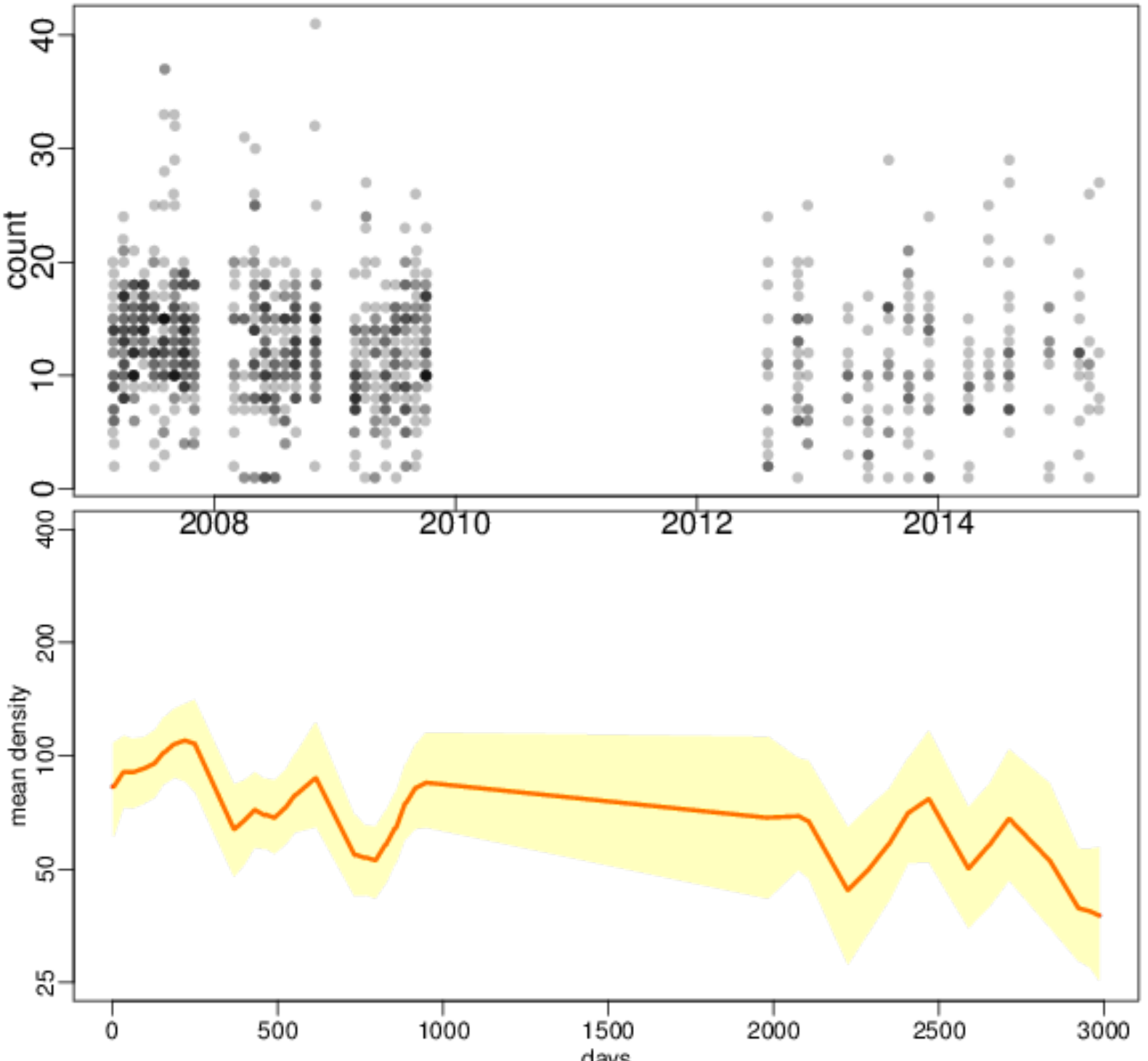
欠測とか不等間隔とか

状態空間モデル + 観測モデル

欠測があっても問題ない
「補完」の必要なし!



不等間隔データでも何とかなります!

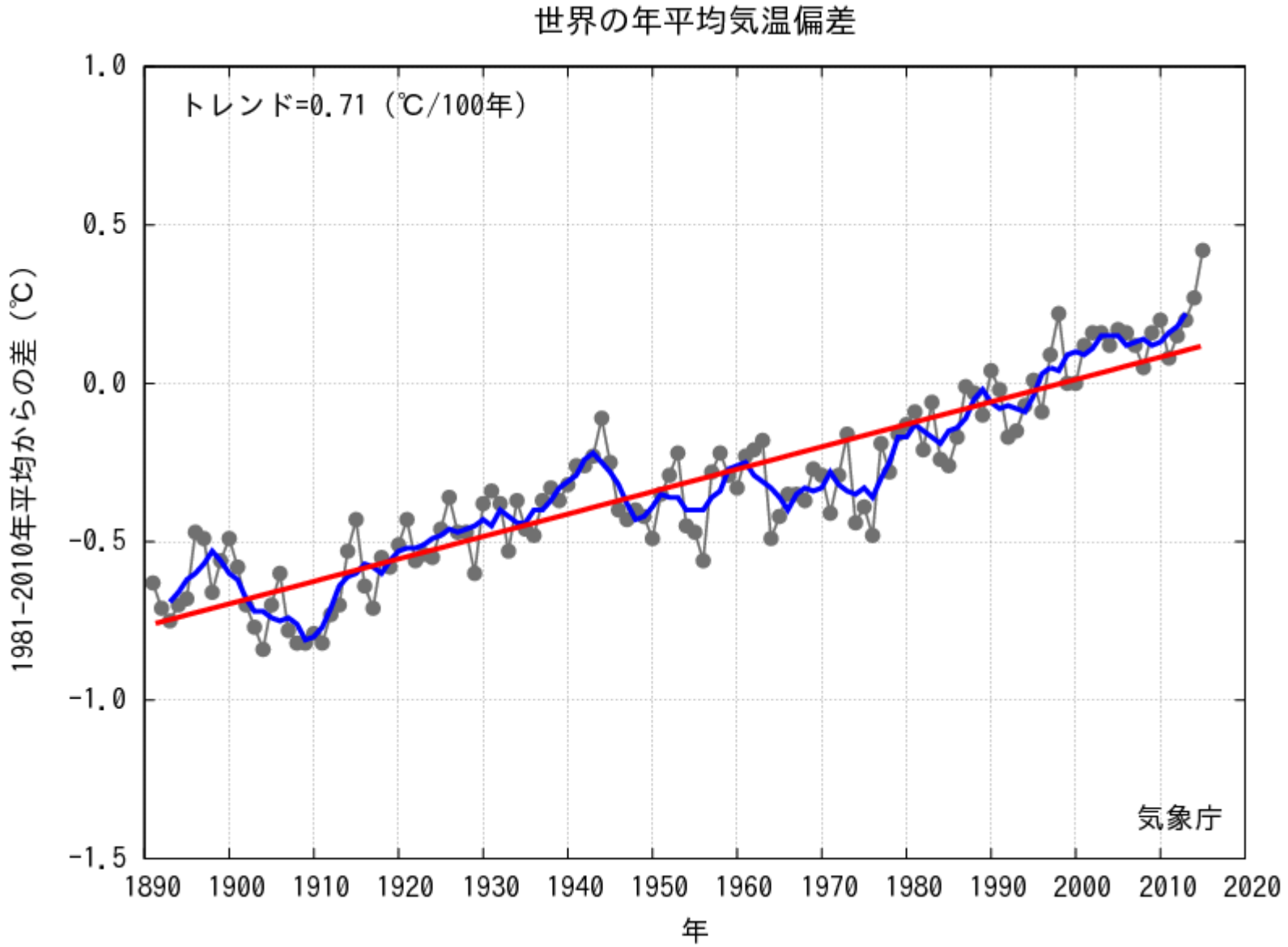


状態空間モデルを使う利点

「ばらばら解析」の回避

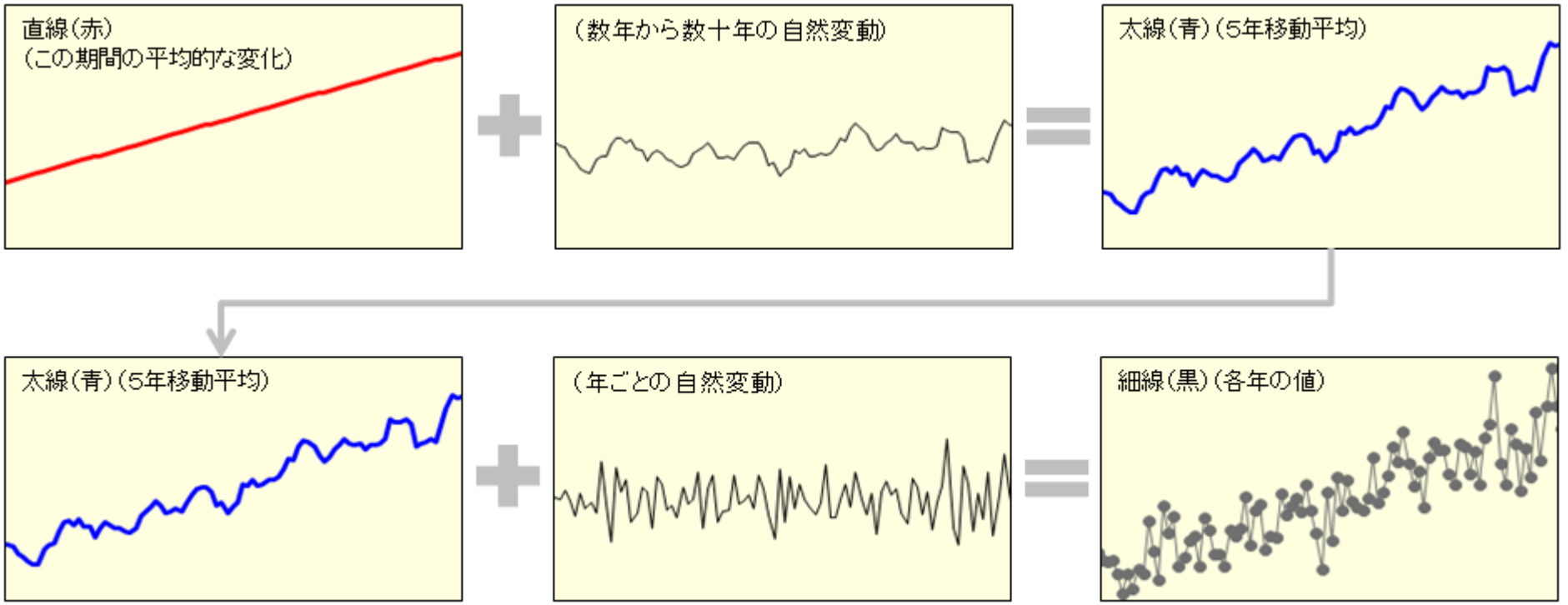
気象庁のデータ解析？

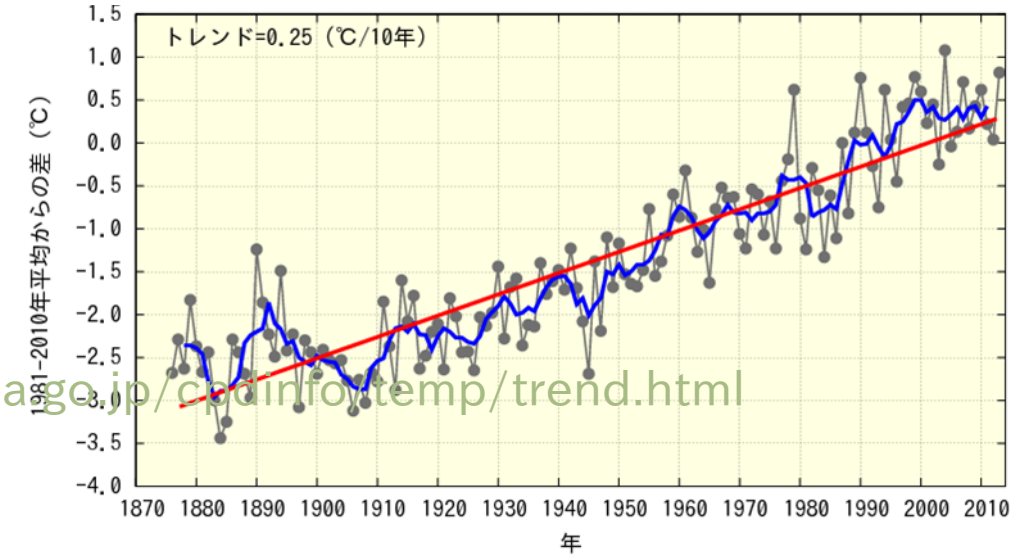
気象庁の長期変化傾向（トレンド）の解説



http://www.data.jma.go.jp/cpdinfo/temp/an_wld.html

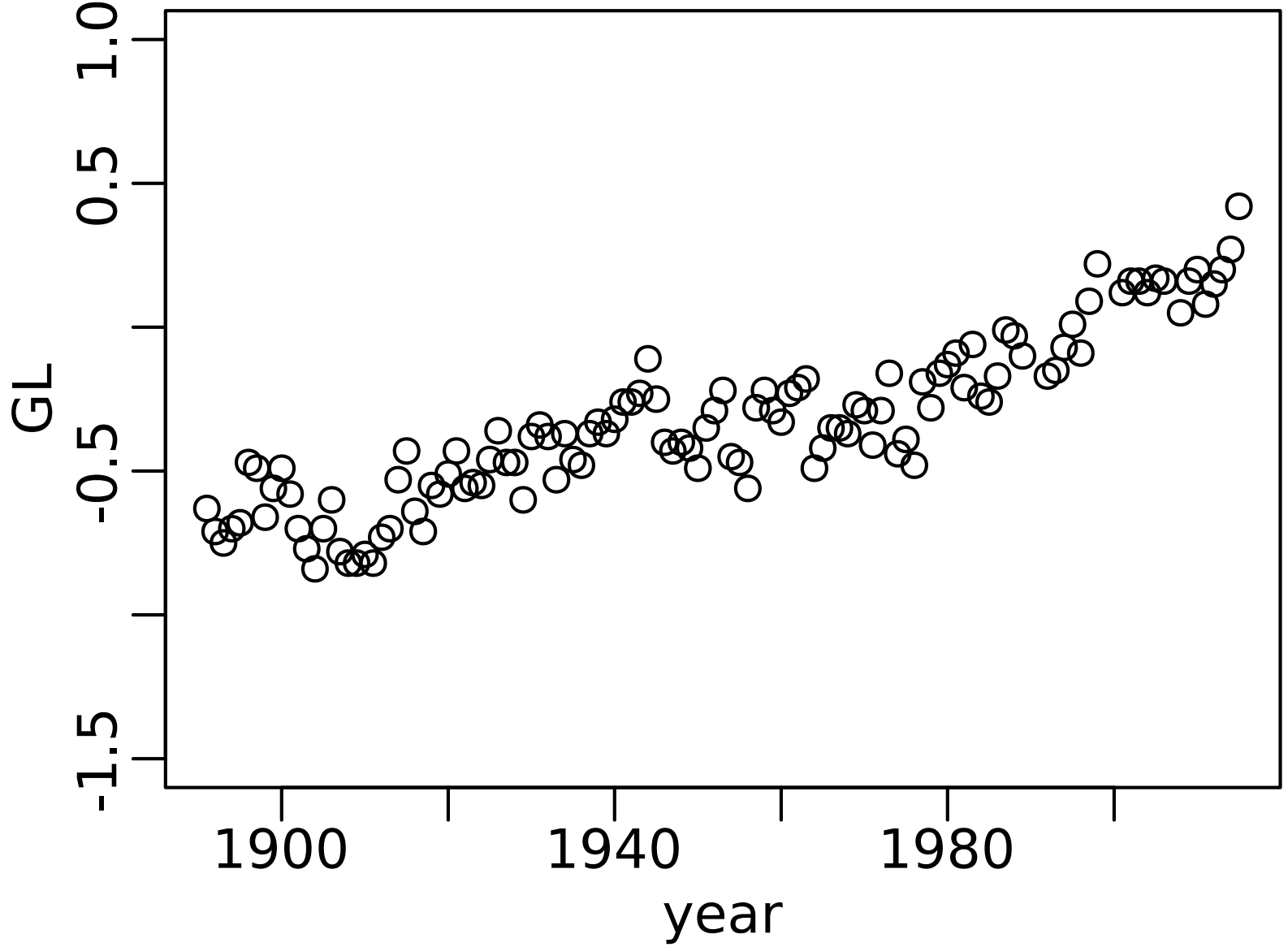
気象庁の長期変化傾向（トレンド）の解説





http://www.data.jmago.jp/cpdinfo/temp/trend.html

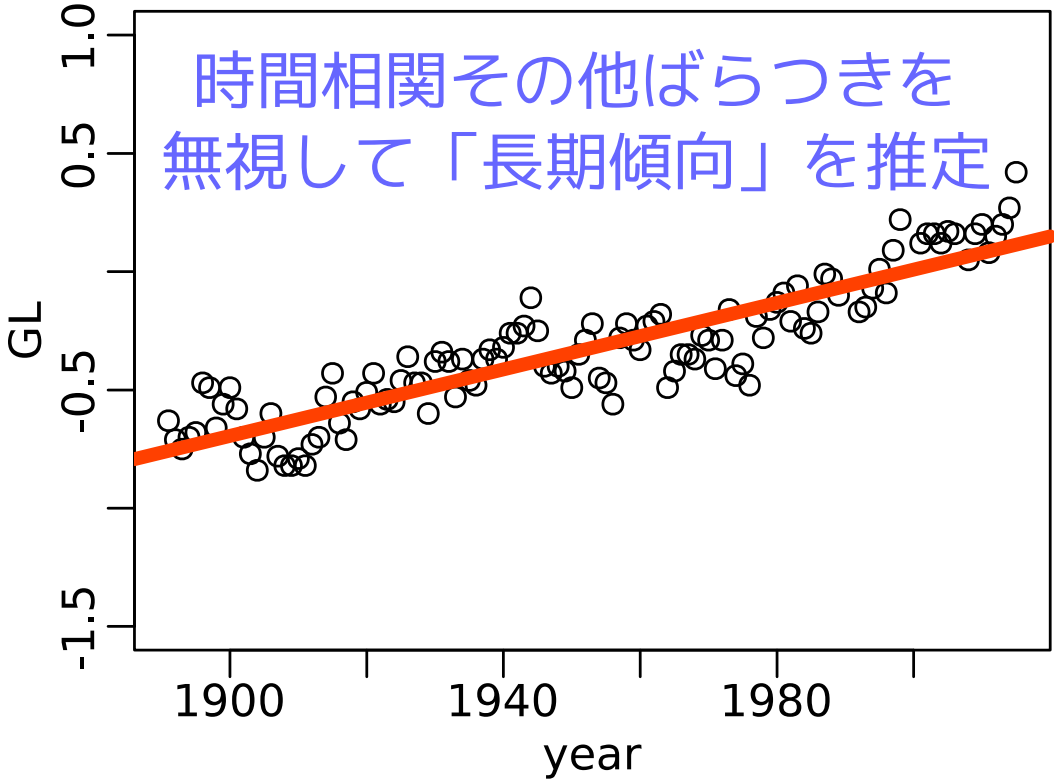
公開データをダウンロード



「とりあえず，直線回帰」の危険性

```
> summary(glm(GL ~ year, data = d))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.41e+01   6.21e-01   -22.6   <2e-16
year         7.03e-03   3.18e-04    22.1   <2e-16
```

100年あたり0.70℃

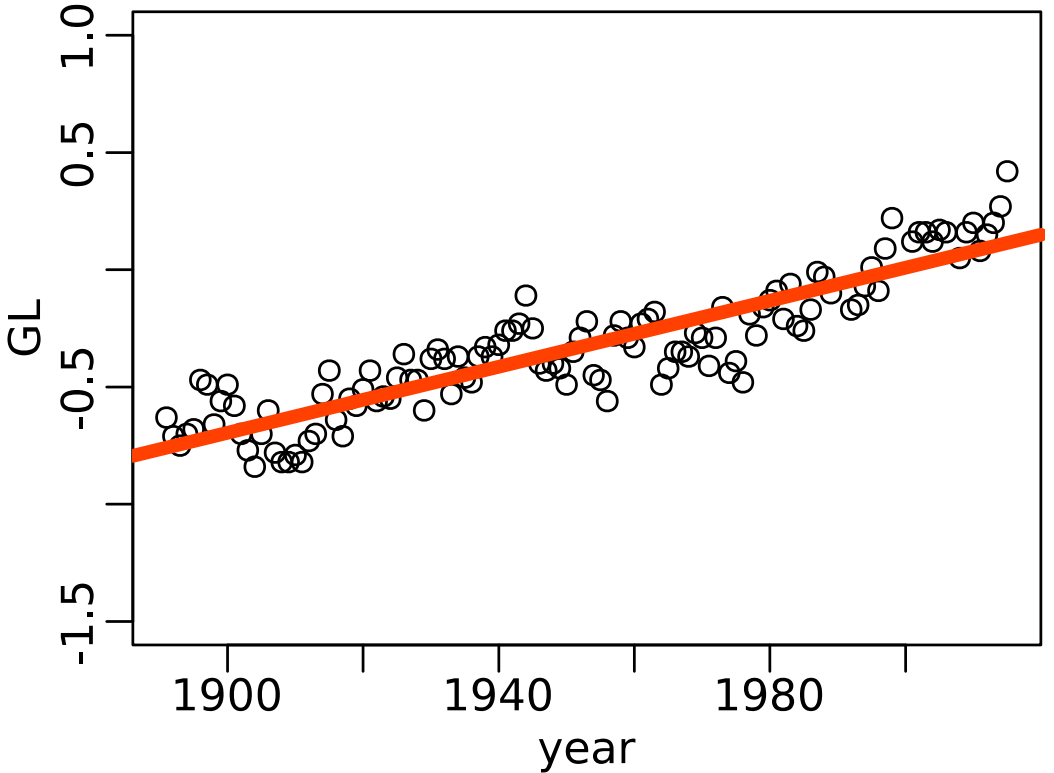


確率 1京ぶんの 2?

直線あてはめ（GLM）が予測した「温暖化」

```
> summary(glm(GL ~ year, data = d))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.41e+01   6.21e-01   -22.6   <2e-16
year         7.03e-03   3.18e-04    22.1   <2e-16
```

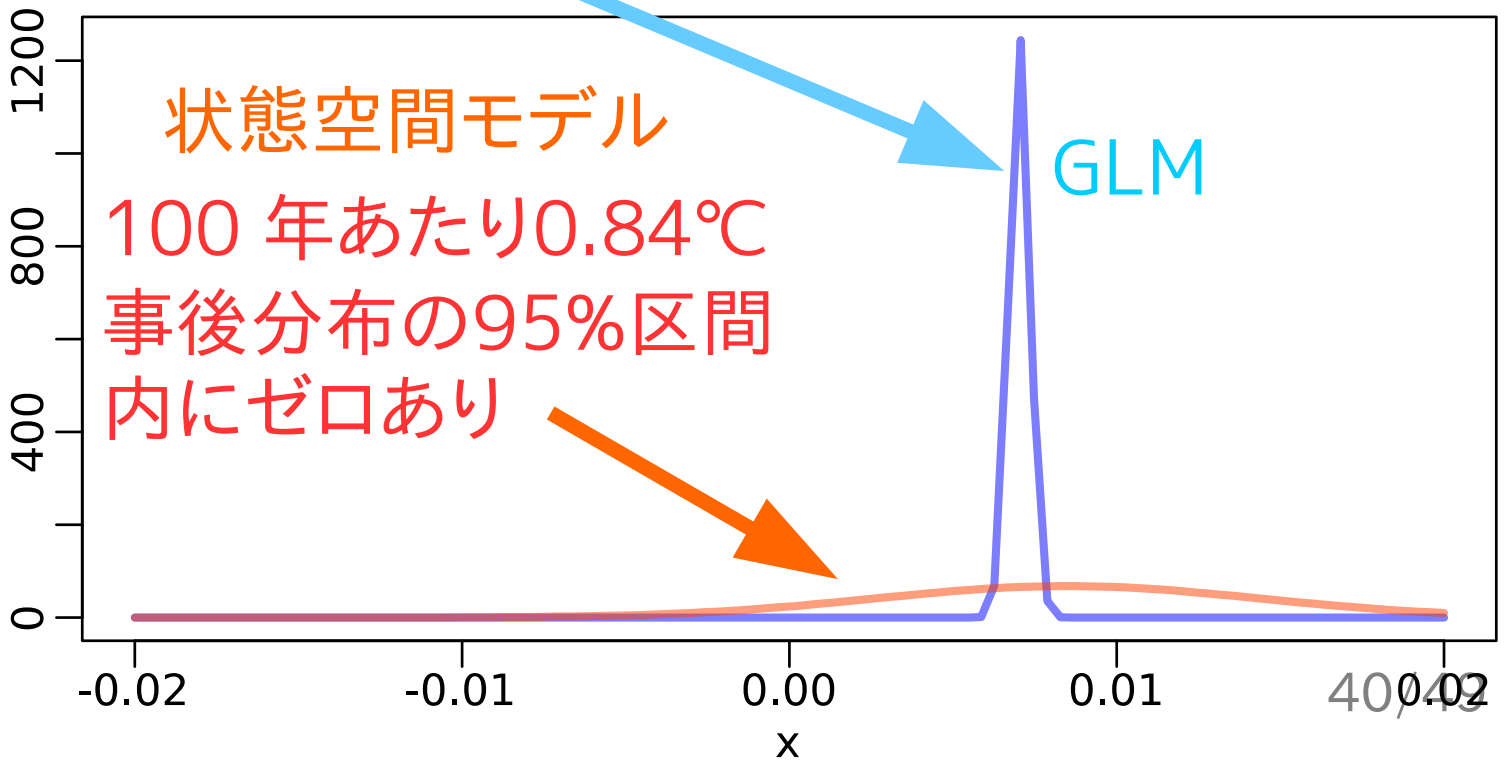
100年あたり0.70℃



状態空間モデルが予測した「温暖化」

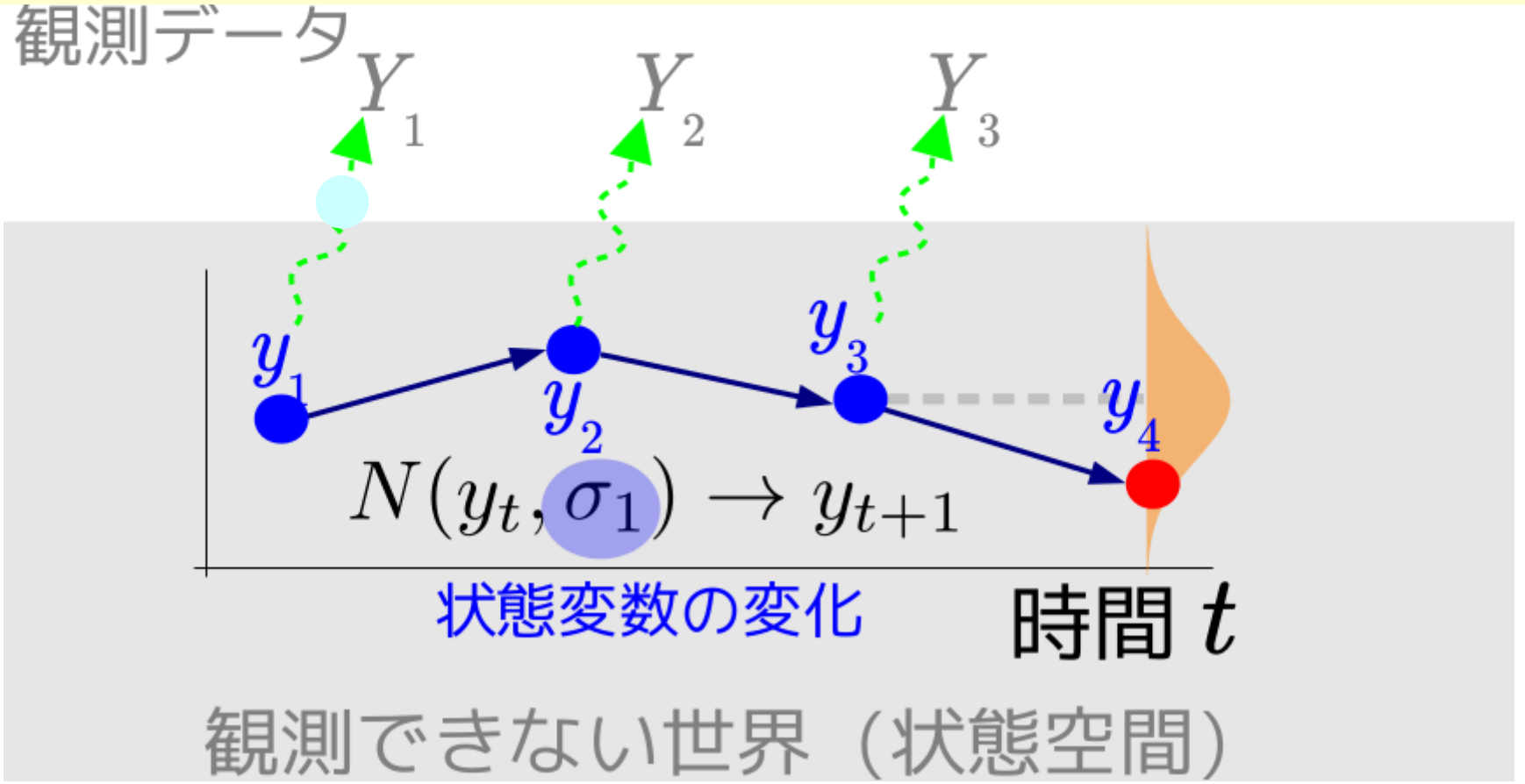
```
> summary(glm(GL ~ year, data = d))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.41e+01   6.21e-01   -22.6   <2e-16
year         7.03e-03   3.18e-04    22.1   <2e-16
```

100年あたり0.70℃



状態空間モデル：すべてを同時に推定

ランダムウォーク+各年独立なノイズ



短い時系列データ

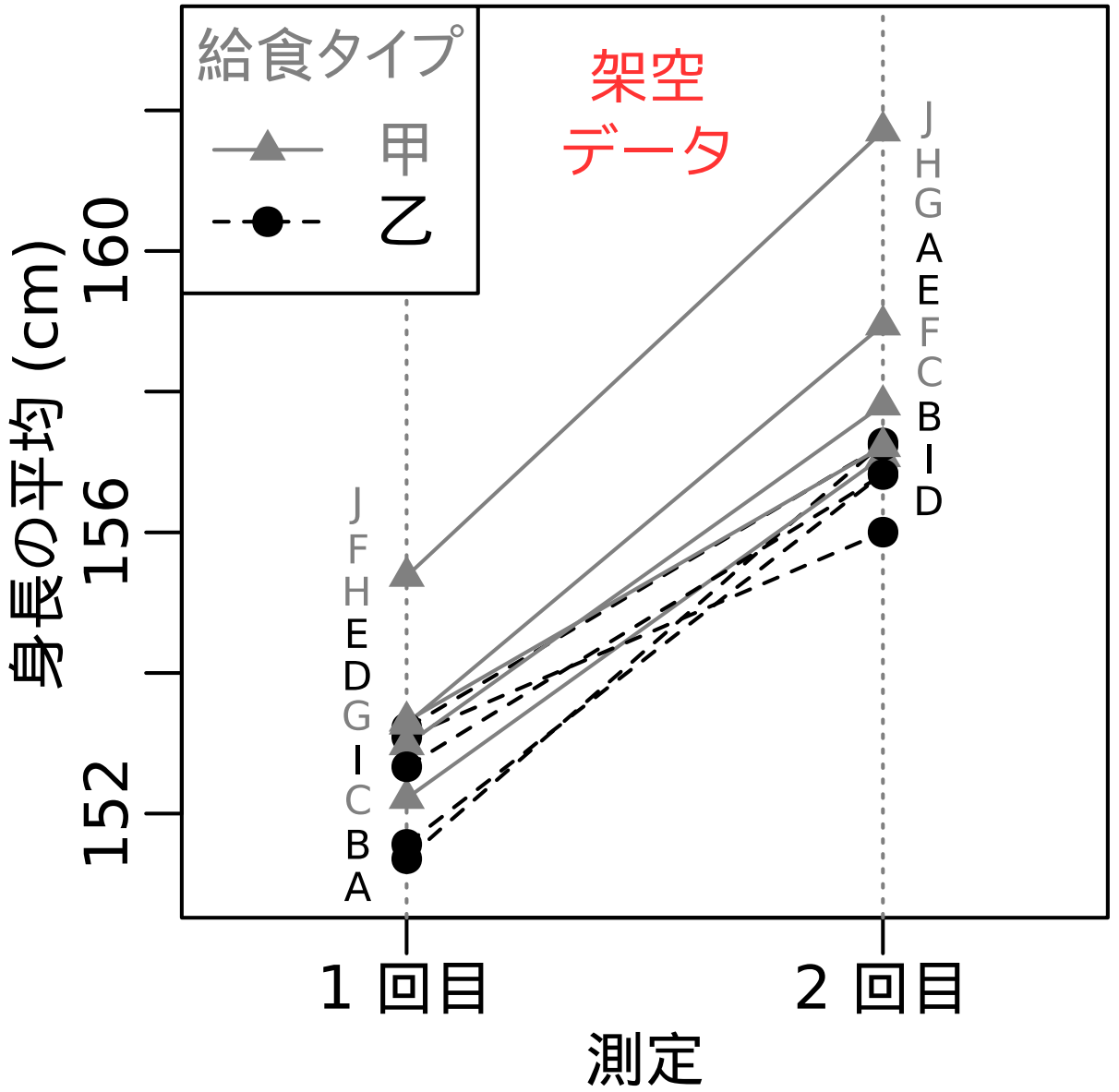
時系列の長短に関係なく

「対応のある」データ点か

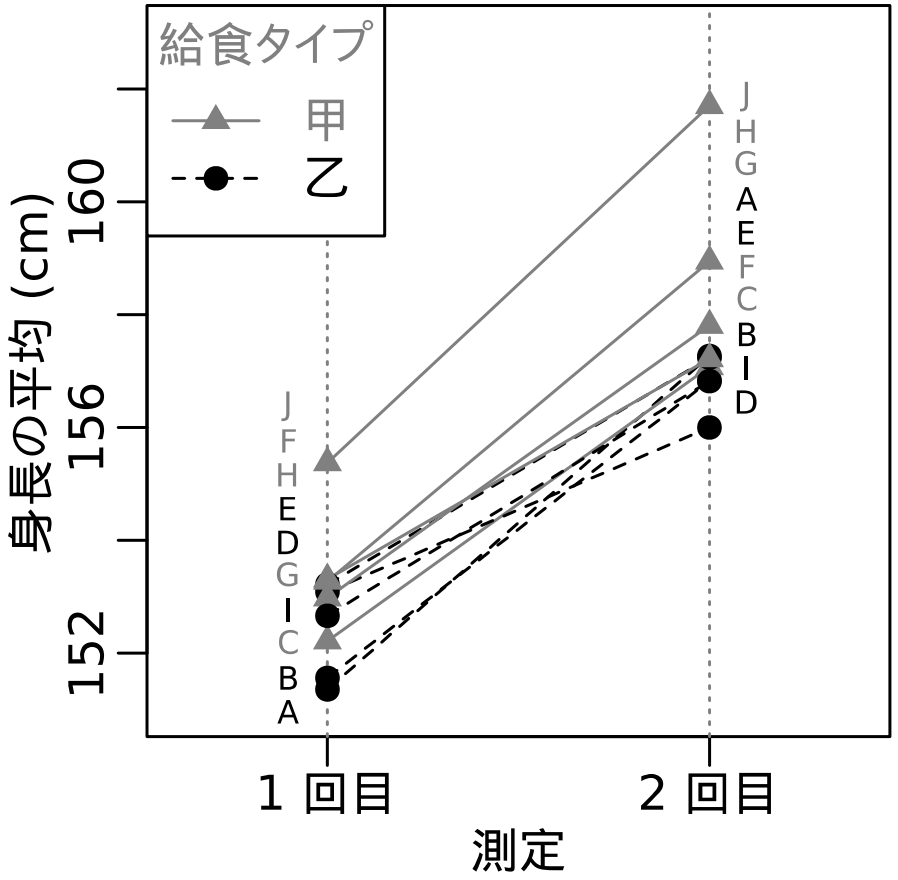
どうかが本質的な問題

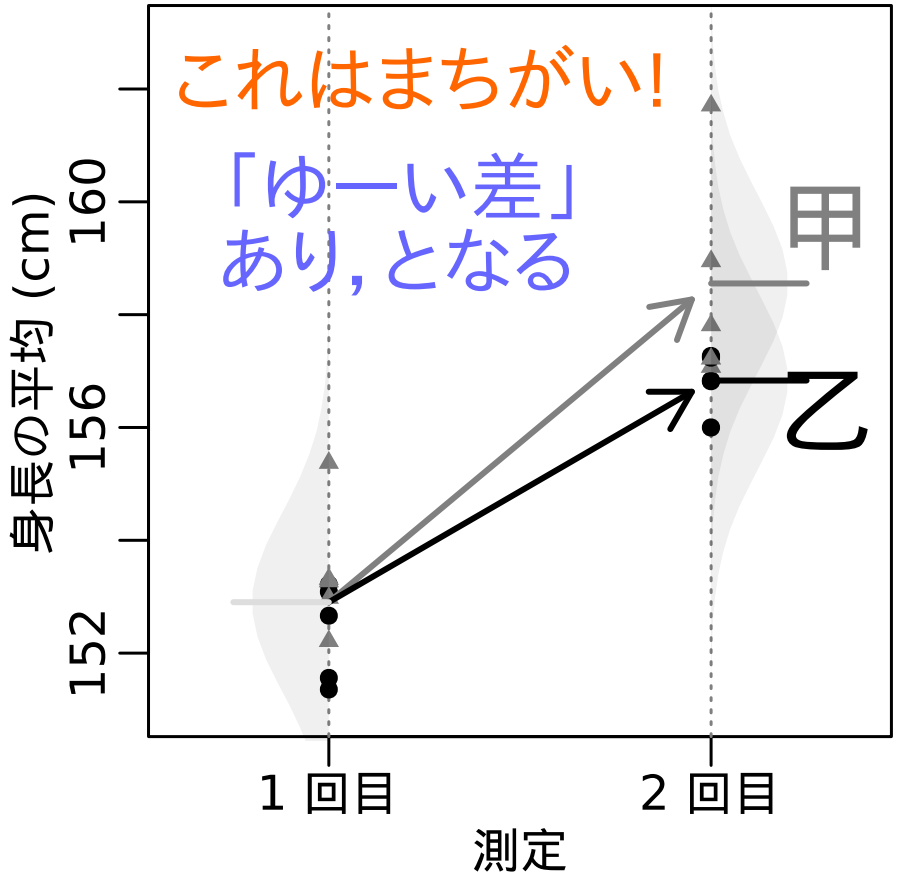
再測定もまた時系列データ

岩波データサイエンス vol.1



対応（paired）を考えてない GLM あてはめ

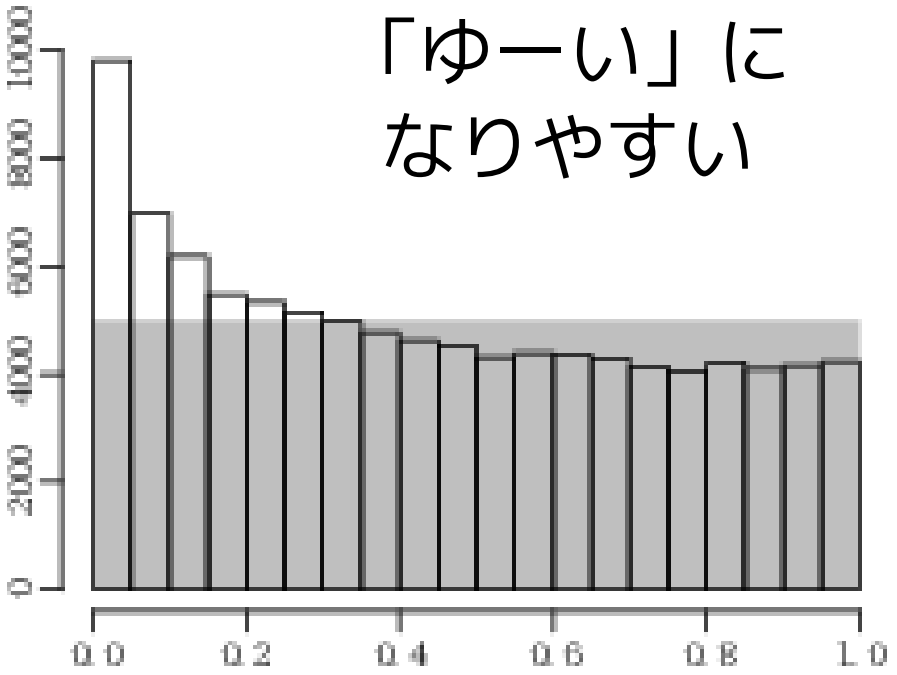




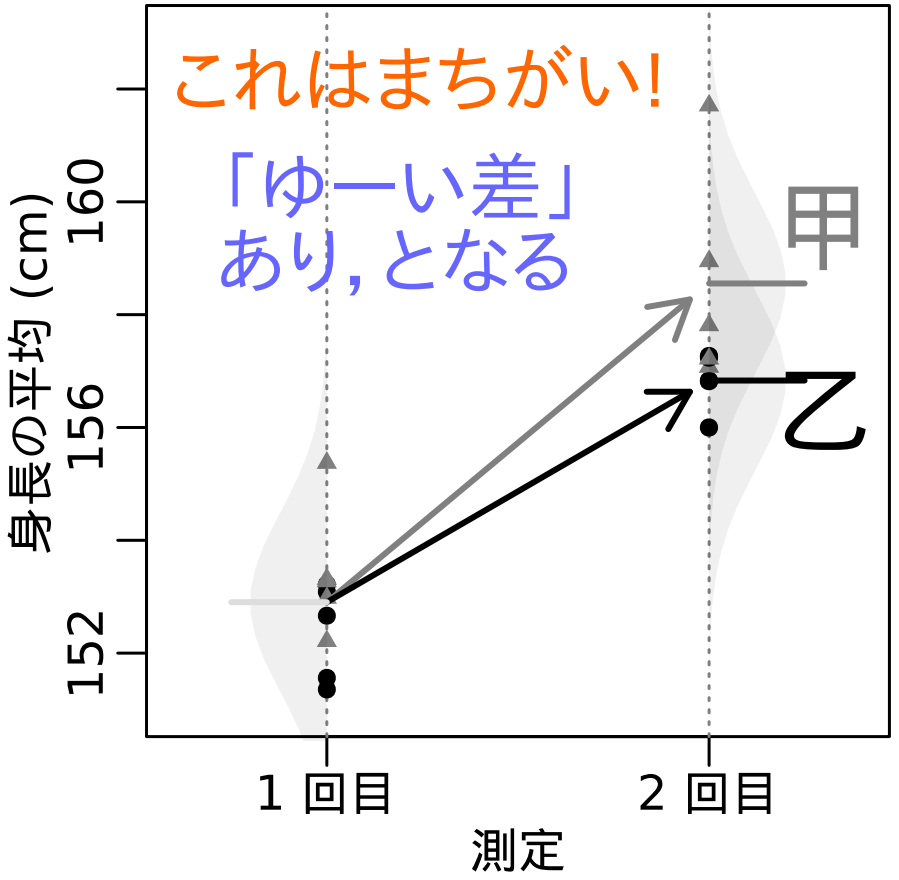
glm(身長 ～ (測定2回目) + (測定2回目):(処理の効果))

同じ対象を二回測定していることを考慮してない

対応（paired）を考えてない GLM あてはめ



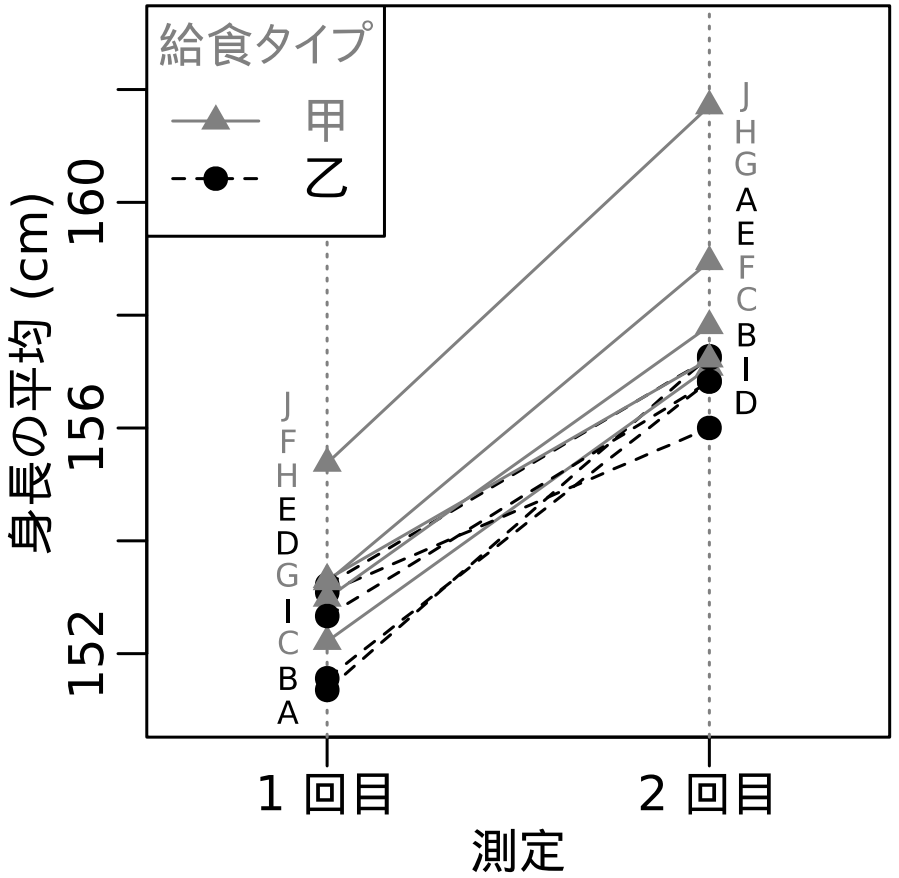
「ゆーい」に
なりやすい

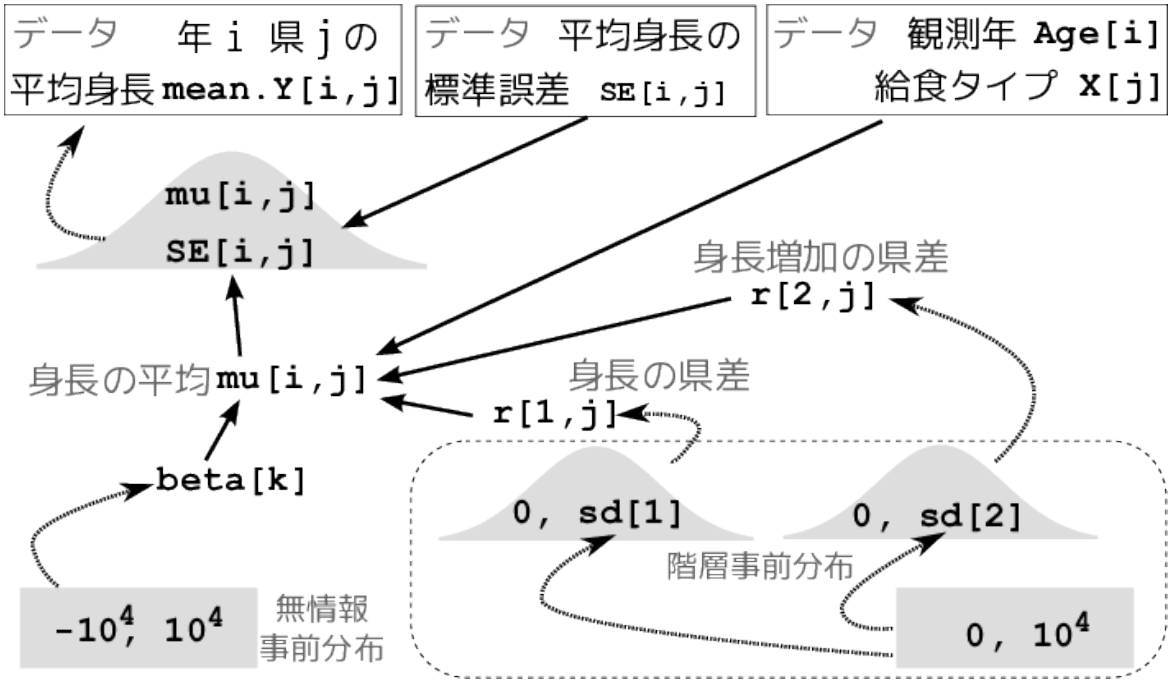


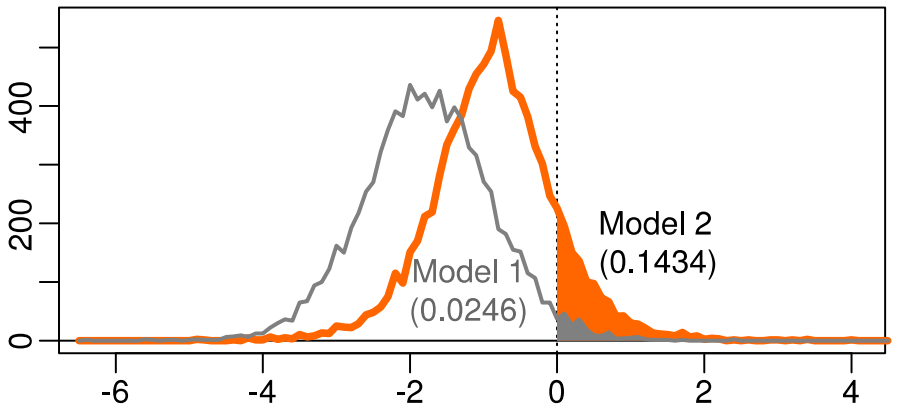
glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))

同じ対象を二回測定していることを考慮してない

対応（paired）を考慮し，さらに県の差もあるモデル







給食効果なし

おわりに

時系列データの統計モデリング でやめたほうがいいこと

- GLM: Y(t) ~ t とか Y(t) ~ X(t)
- 段階的解析:観測値の四則演算
- 「残差」の再解析
- 「対応」の無視 – 再測は時系列

今回，説明してみたこと

- 時系列データ:単純な回帰はダメ(続)
- 状態空間モデル: 乱歩と雑音の分離
- 差分と時間的自己相関係数
- 欠測と不等間隔
- 時系列と「対応のある」データ
- 説明しないこと - 因果推定など