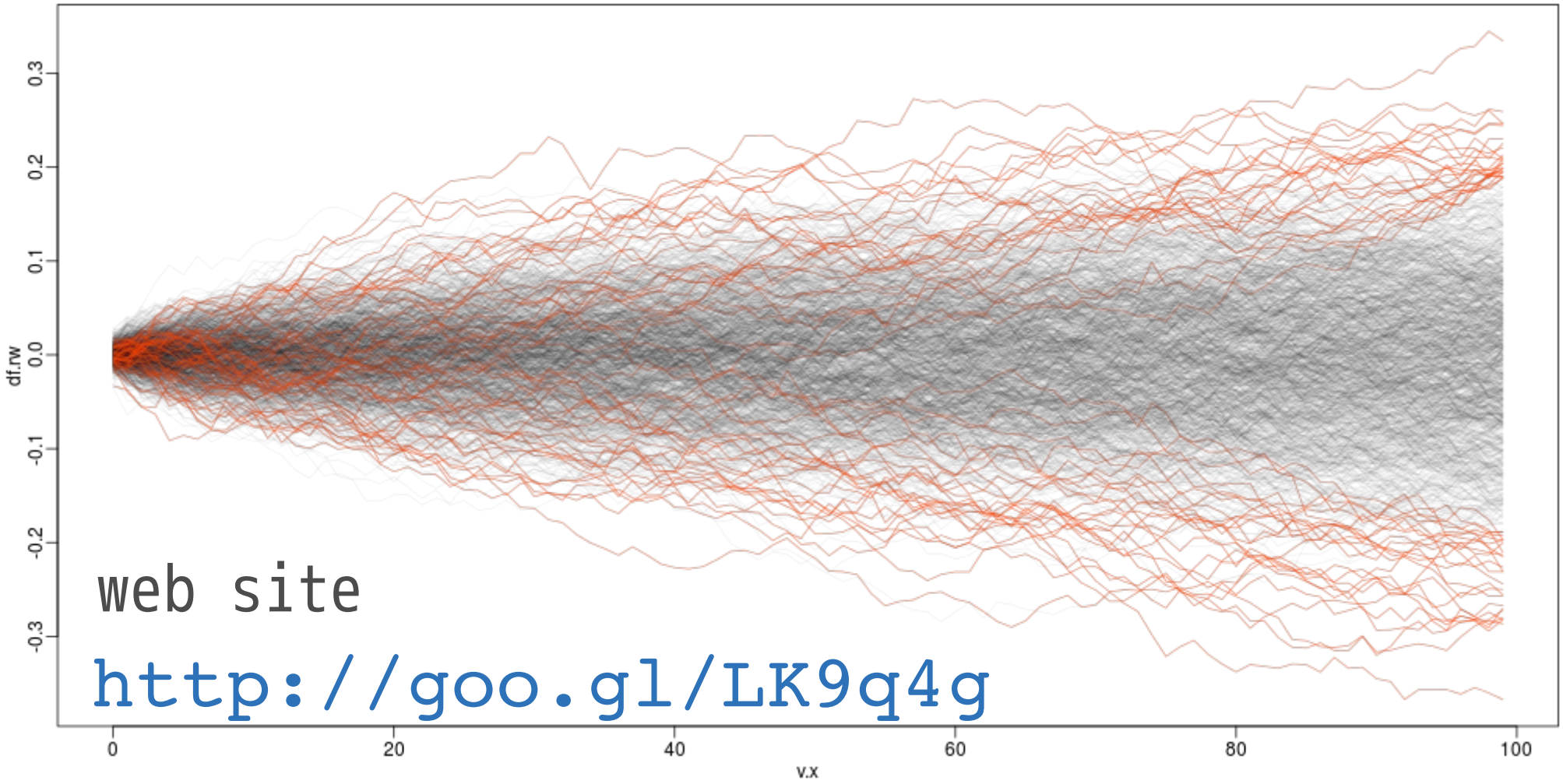
時系列データの統計モデリング入門

久保拓弥（北海道大・環境科学）

kubo@ees.hokudai.ac.jp　@KuboBook



web site

http://goo.gl/LK9q4g

自己紹介：久保拓弥

- 北大の環境科学院という学部のない大学院
- 生態学に関係するデータ解析とかやってます
  - 野外調査をしない生態学者
- データは誰か別の人（共同研究者）がとってきてくれます

そもそも生態学って何?

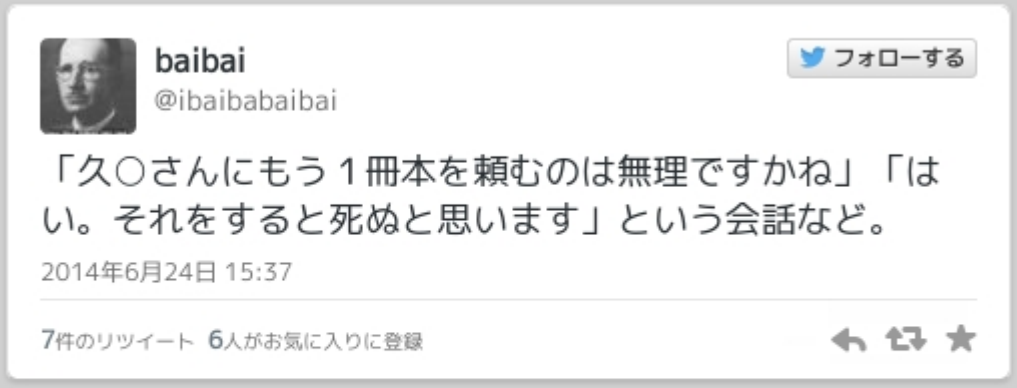
- 生物の数の変化や分布や生活の様子を調べる
- いろいろな動植物が対象

「あれするな」「これやれ」という
おしつけが多い本?

- そもそも R つかうしかないでしょ?
- なんでも正規分布，はだめ
- 観測値どうしの「わり算」するな
- 検定，そんなにいいか? モデル選択のほうがよくない?
- けっきょく random effects は必須でしょ?
- どうせなら階層ベイズ使えるようになったら?
- 最尤推定法より MCMC がいいのでは
- 信頼区間よりも事後分布が便利じゃない?
- 空間相関，無視すんなよ

近況報告：ここ 2.5年間ほど育児ばてでぼろぼろでした……

- お，なにかいのちびろいしたみたいだ ……



- 1720 研究室発．買いもの．保育所でとら回収．

[くまバス]
…… に手をふるクマ好きバス好きのとら．目撃例は二度目．1755 ごろ北 5 条手稲通りに ホッキョクグマラッピングバス が「回送バス」（円山動物園いき?）として通過しているもよう ……

KuboLog ぎょーむ日誌 2014-06-21

- 1810 帰宅．とら風呂．とらめし．晩飯．とら服の洗濯．

今日のハナシ

「あぶない」時系列データ解析

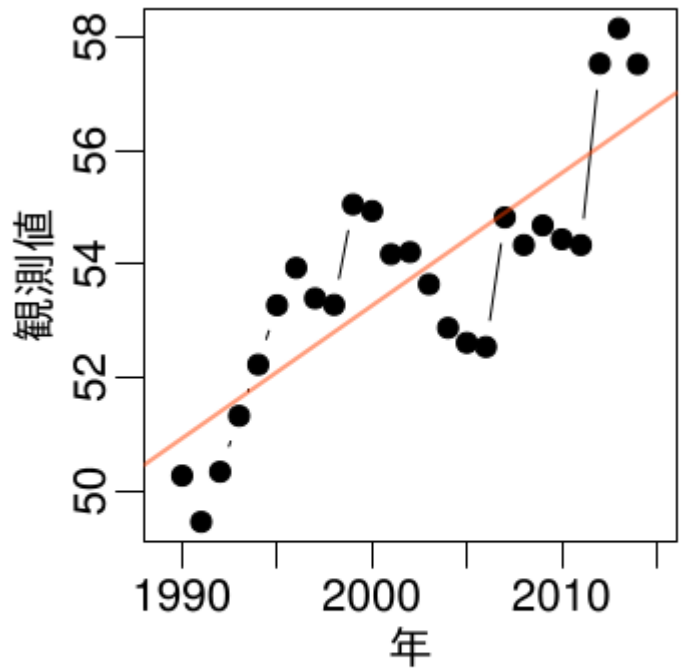
はやめましょう!

(危 1) 時系列データの GLM あてはめ

(危 2) 時系列 Y_t ～ 時系列 X_t

各時刻の個体数 ～ 気温 とか

（危 1）時系列データを GLM で



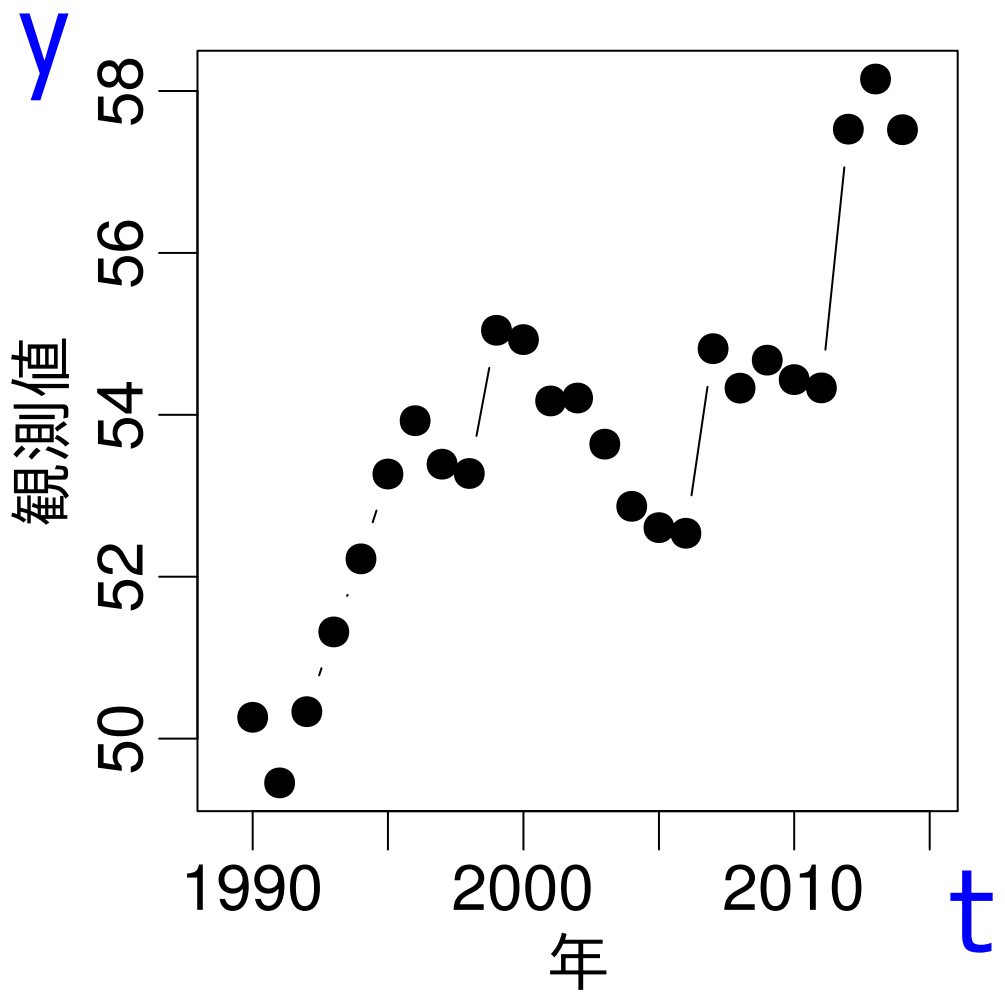
「ゆーいな傾き」を
ねつぞうする原因

傾きの検定やめて
AIC モデル選択
しても同様になる

検定とかモデル選択とかそういう問題ではない

統計モデルがおかしい?

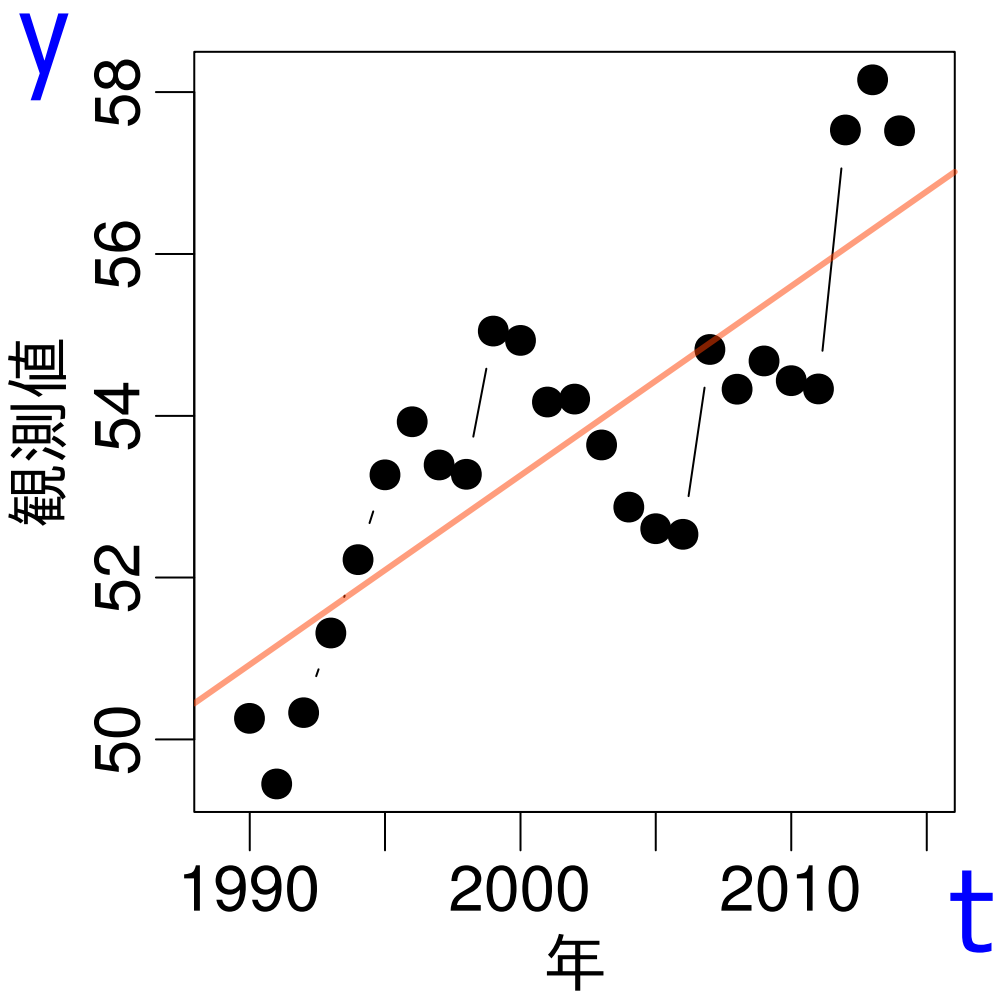
このような時系列データがあったとしましょう



y は何か連続値としましょう

(今日でてくる y は連続値ばかり，ということで)

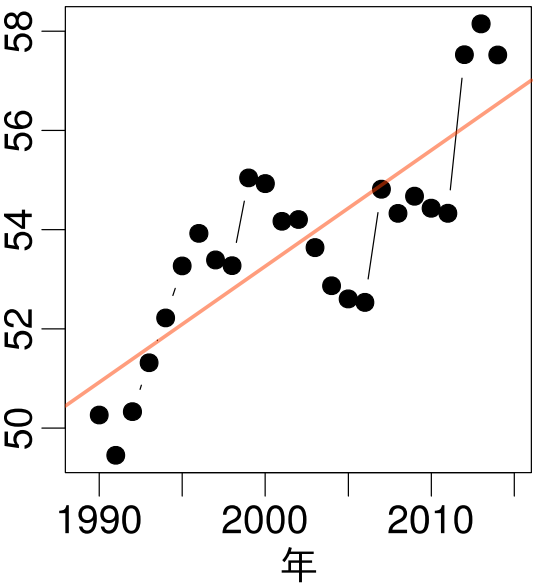
時系列データの統計モデリング入門



glm(y ~ t)

…とモデル
をあてはめてみた

「やったーゆーいだ!!」……??



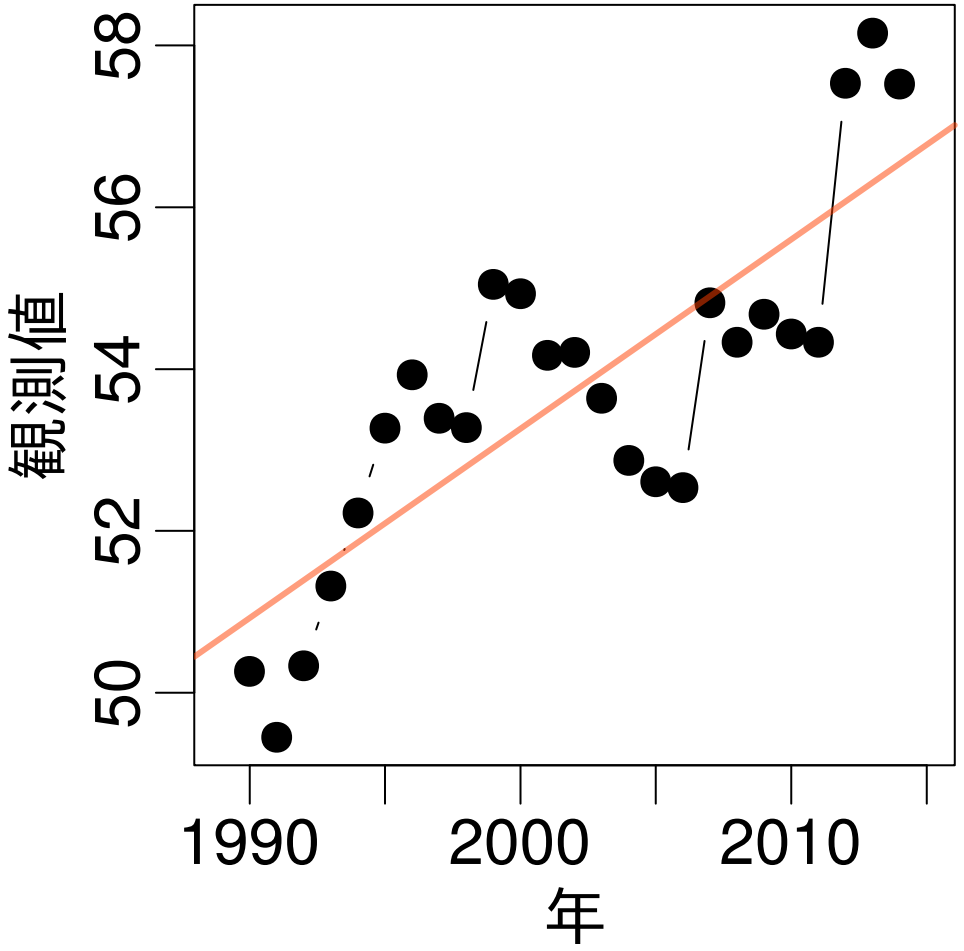

```
> summary(glm(formula = y ~ t))

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)

時系列の各点は独立ではない



「ゆーいな傾き」を

ねつぞうする原因

傾きの検定やめて

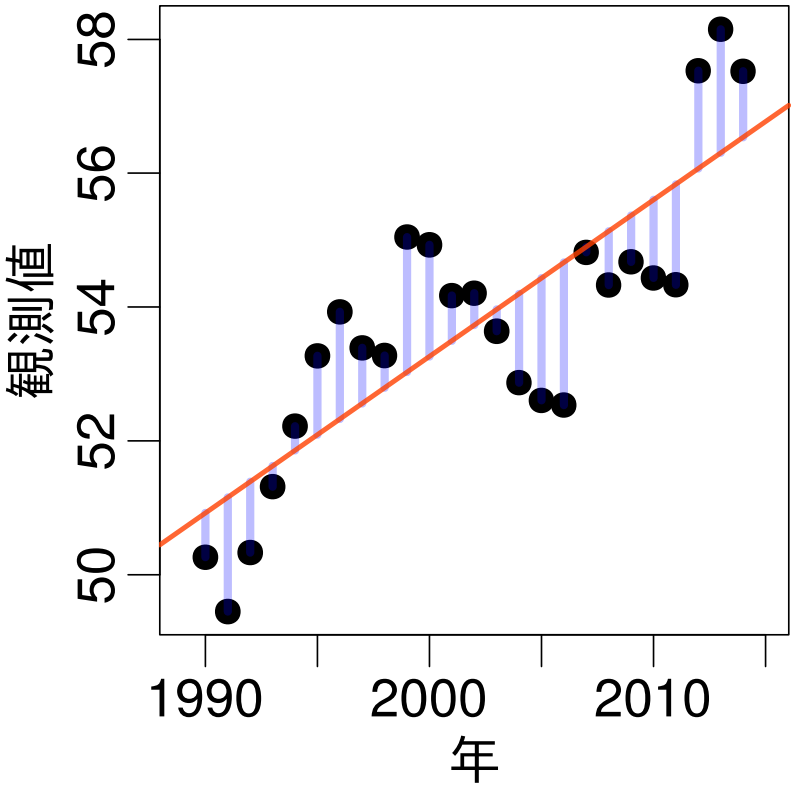
AIC モデル選択

しても同様になる

検定とかモデル選択とかそういう問題ではない

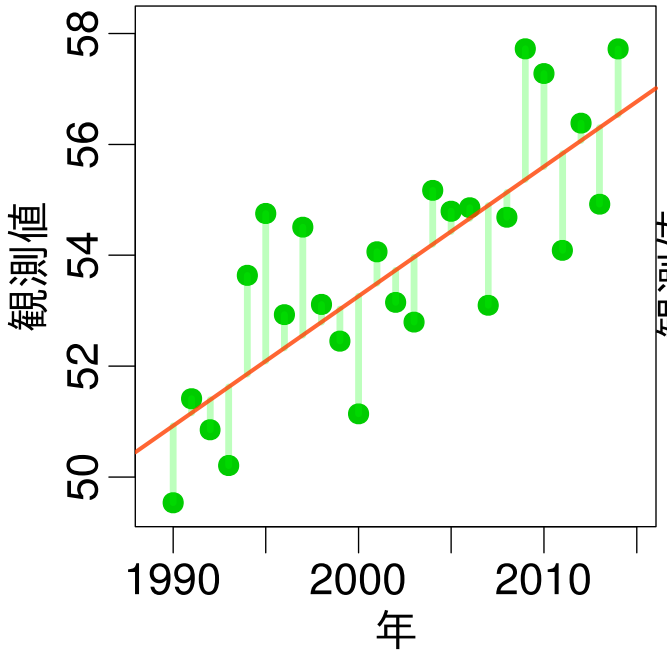
統計モデルがおかしい?

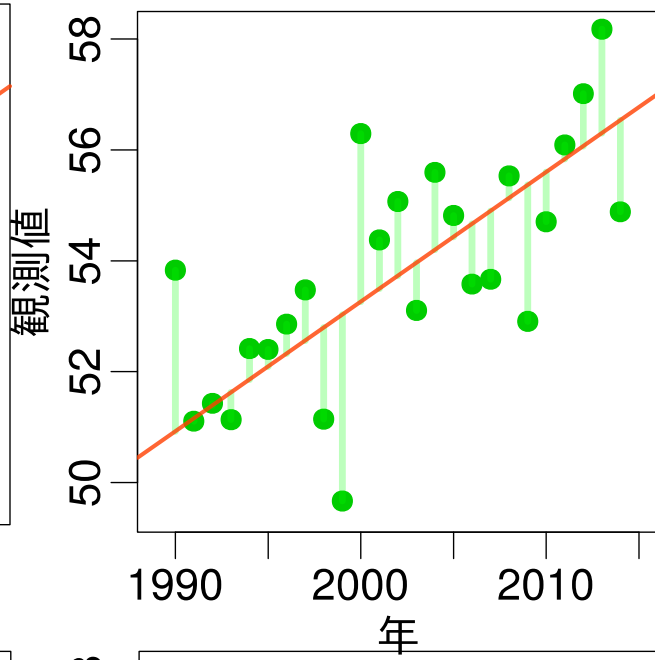
時系列の「ずれ」

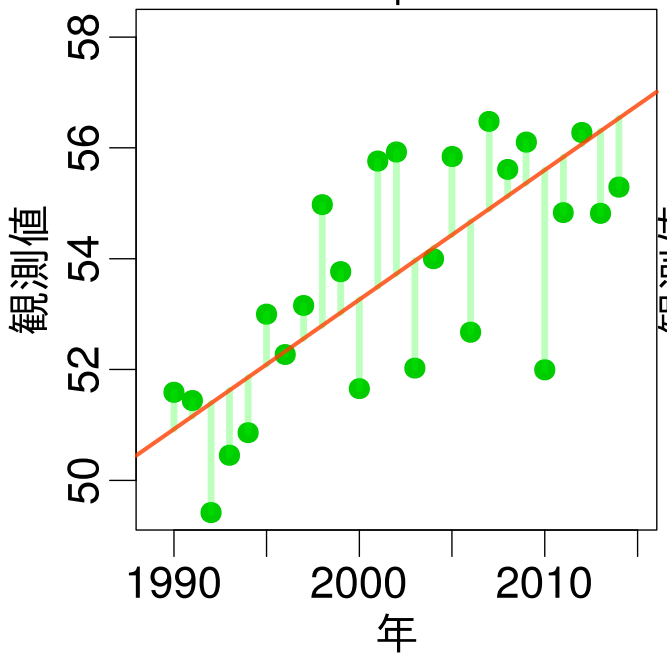


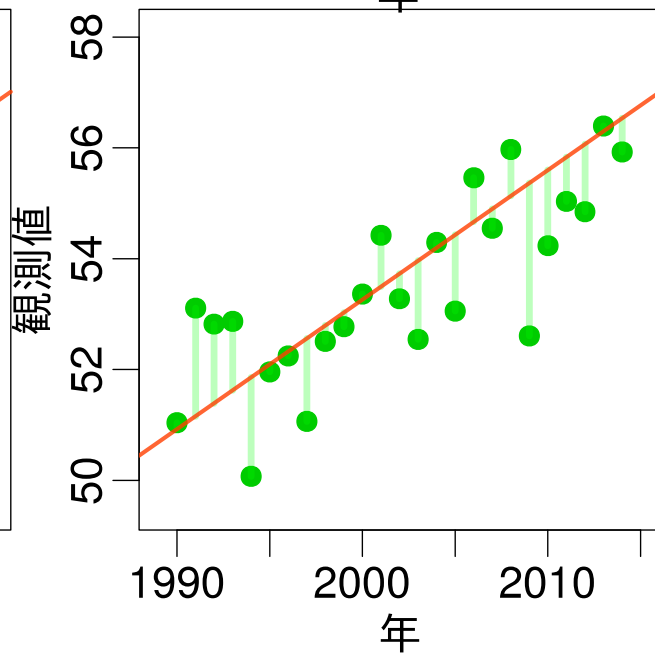
ずれかたが
ちがってる?

GLM のずれ

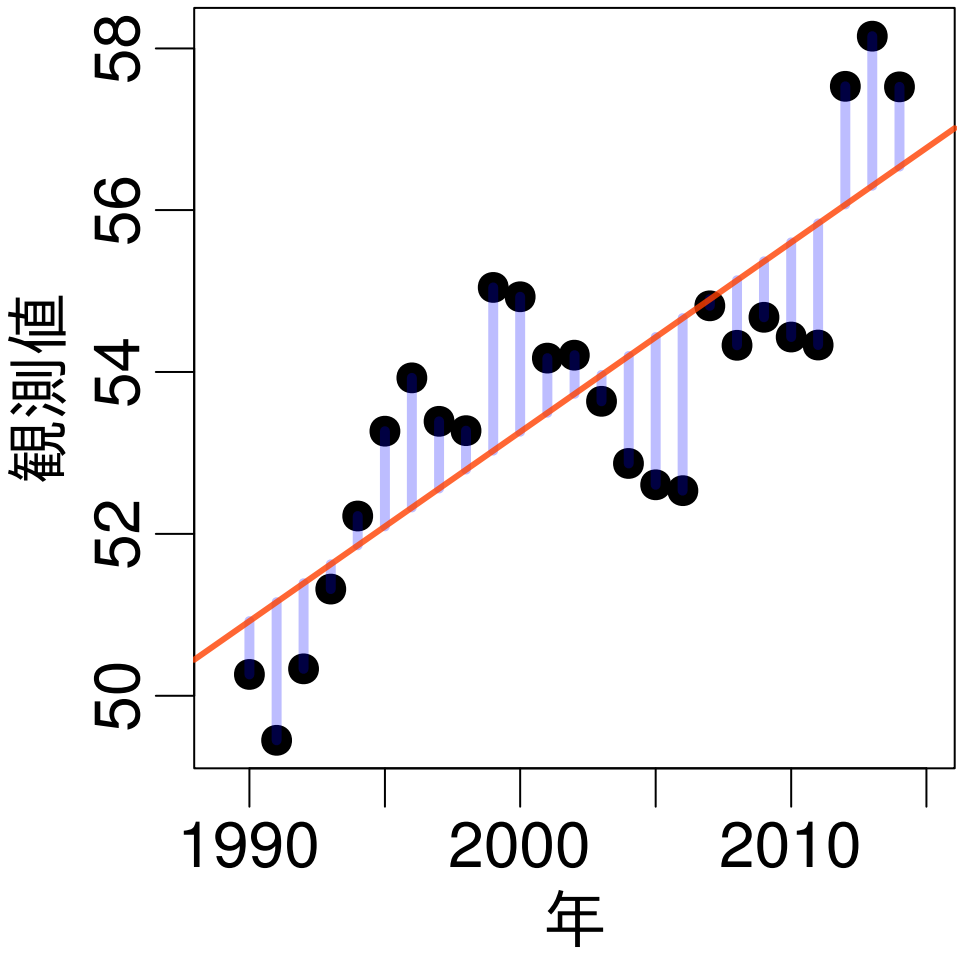




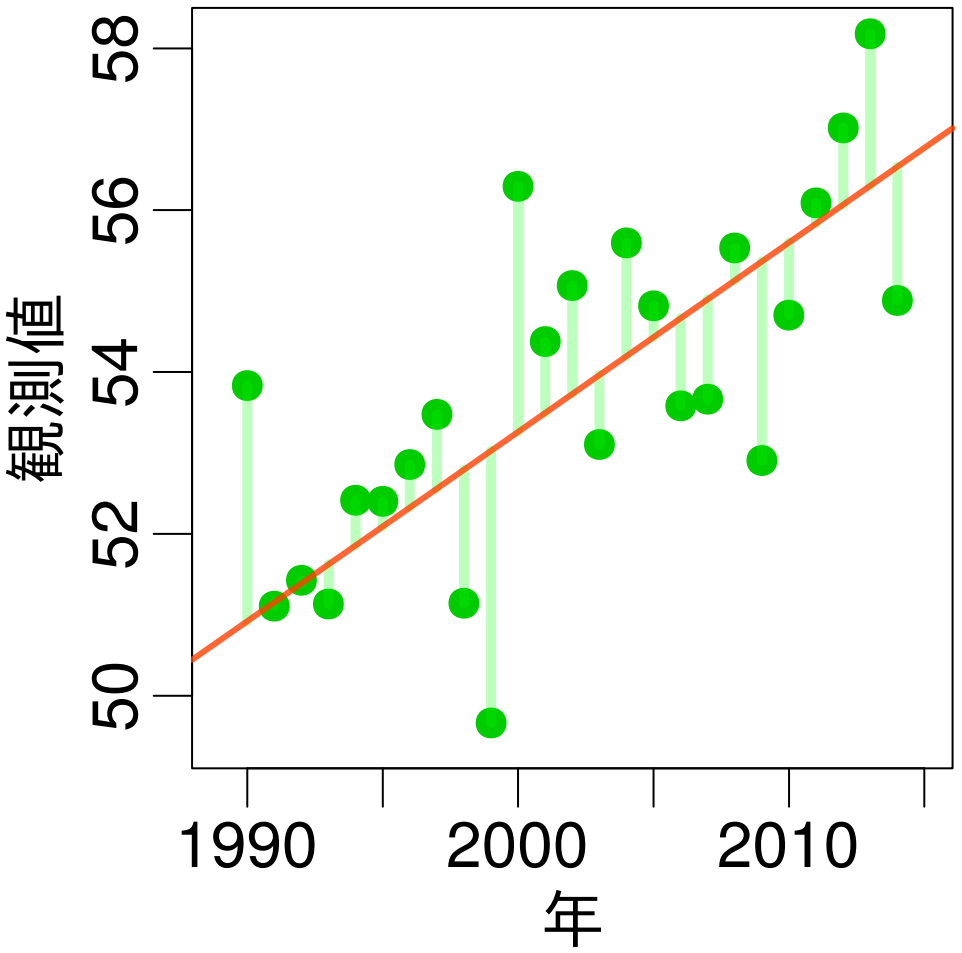




時系列の「ずれ」



GLM のずれ



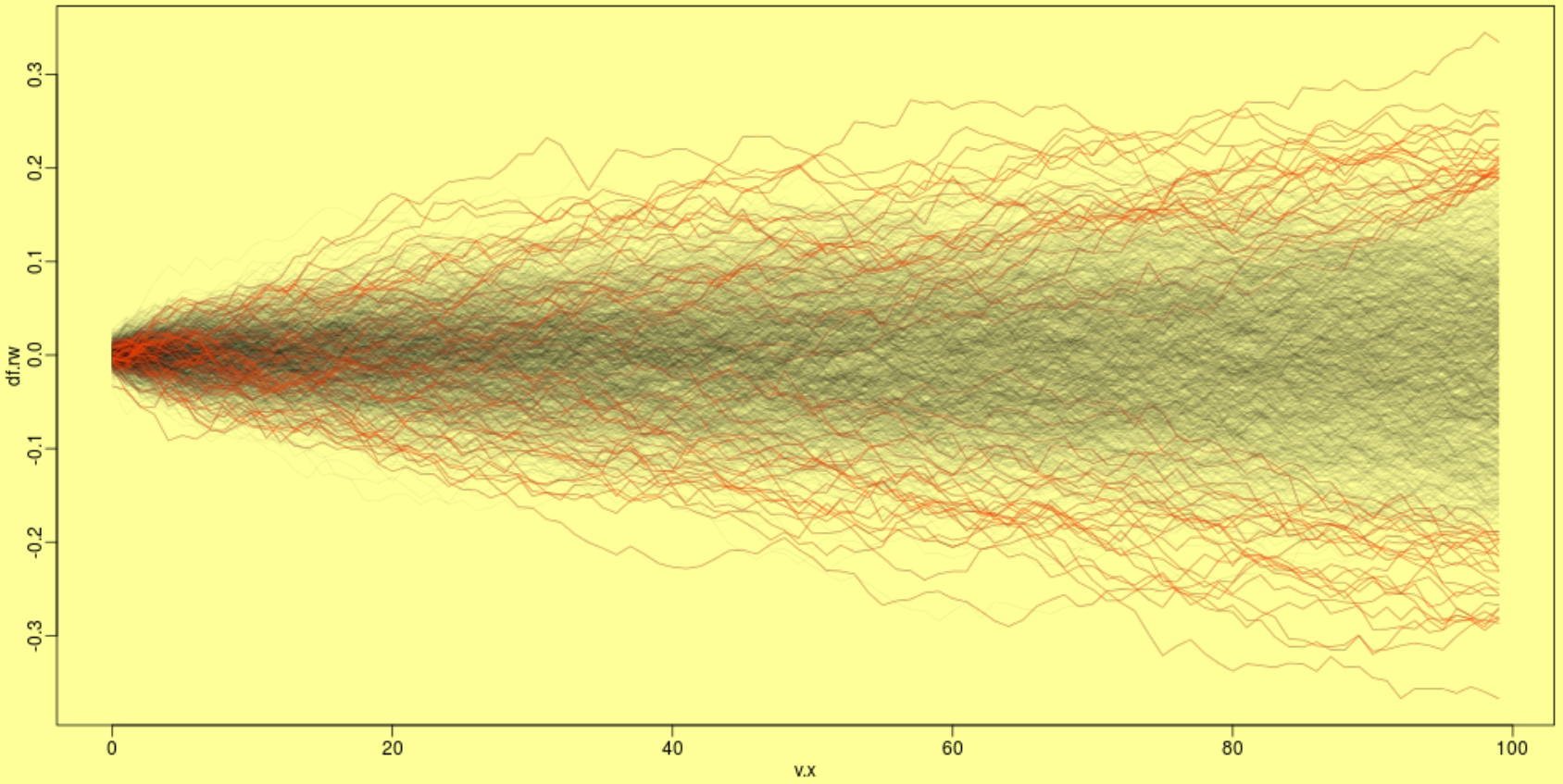
直線からのずれがちがう!

時間的自己相関がある

時間的自己相関がない

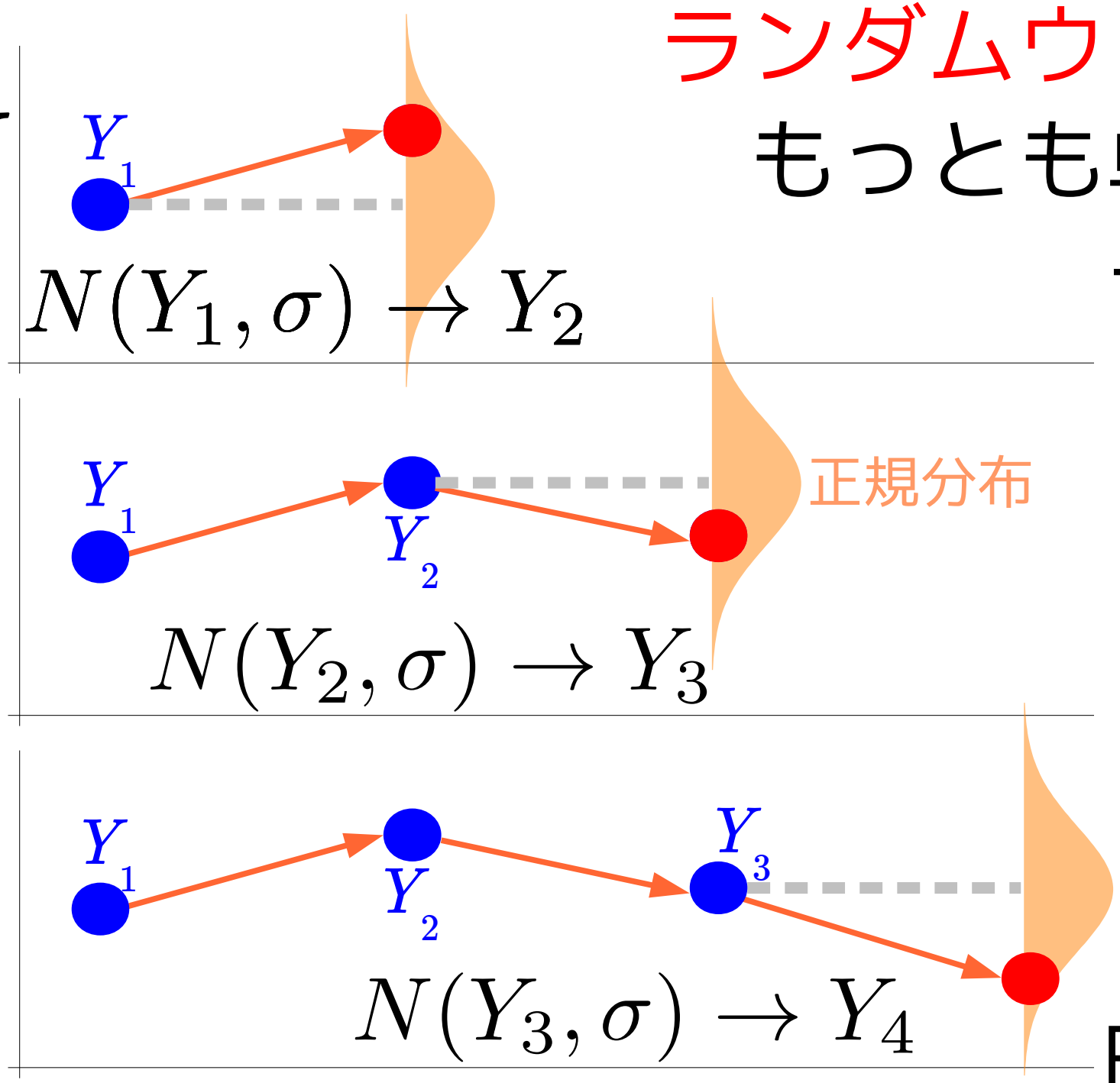
時系列の基本モデルのひとつ

ランダムウォーク（乱歩）



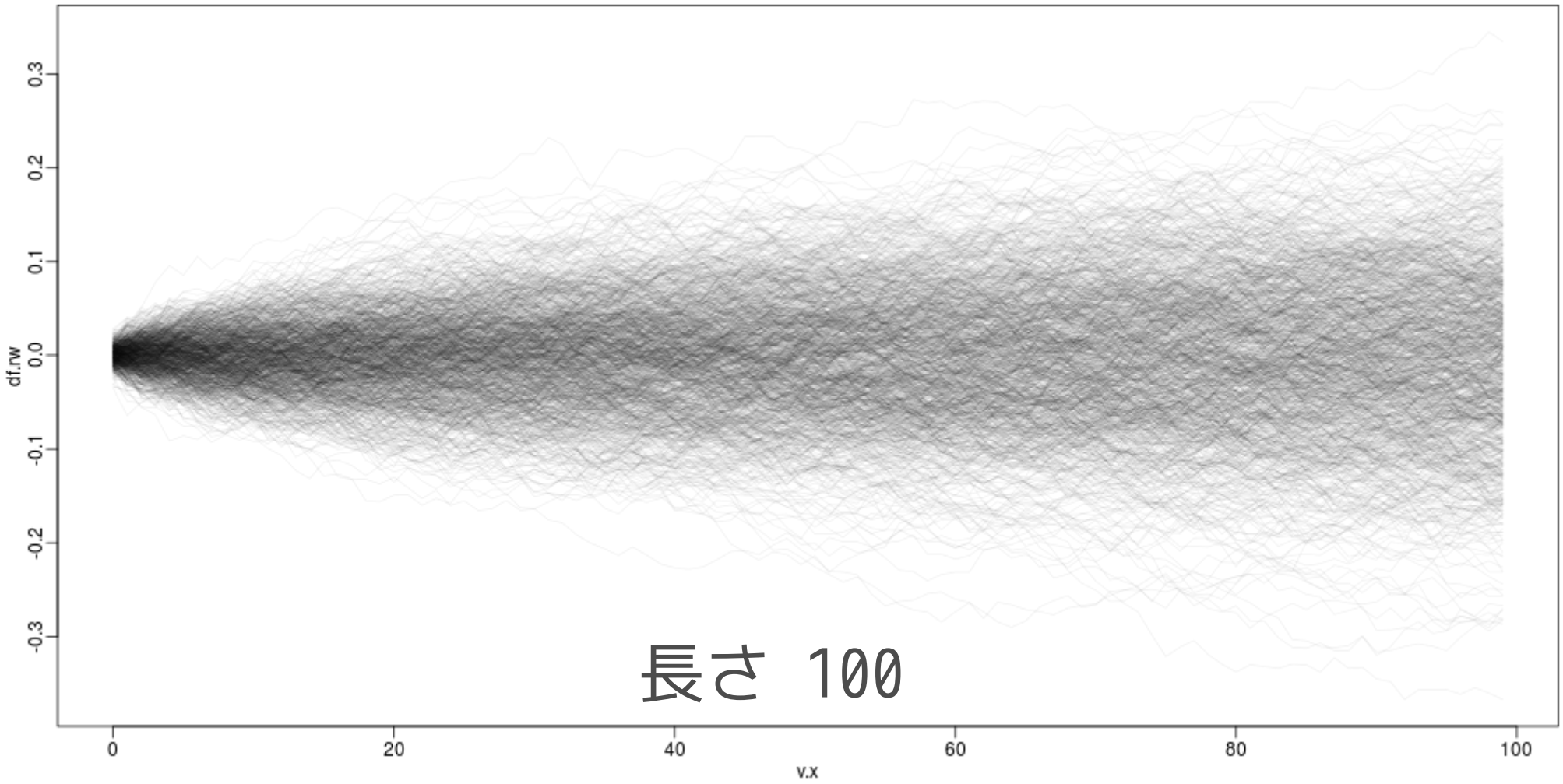
ランダムウォーク

もっとも単純なモデル



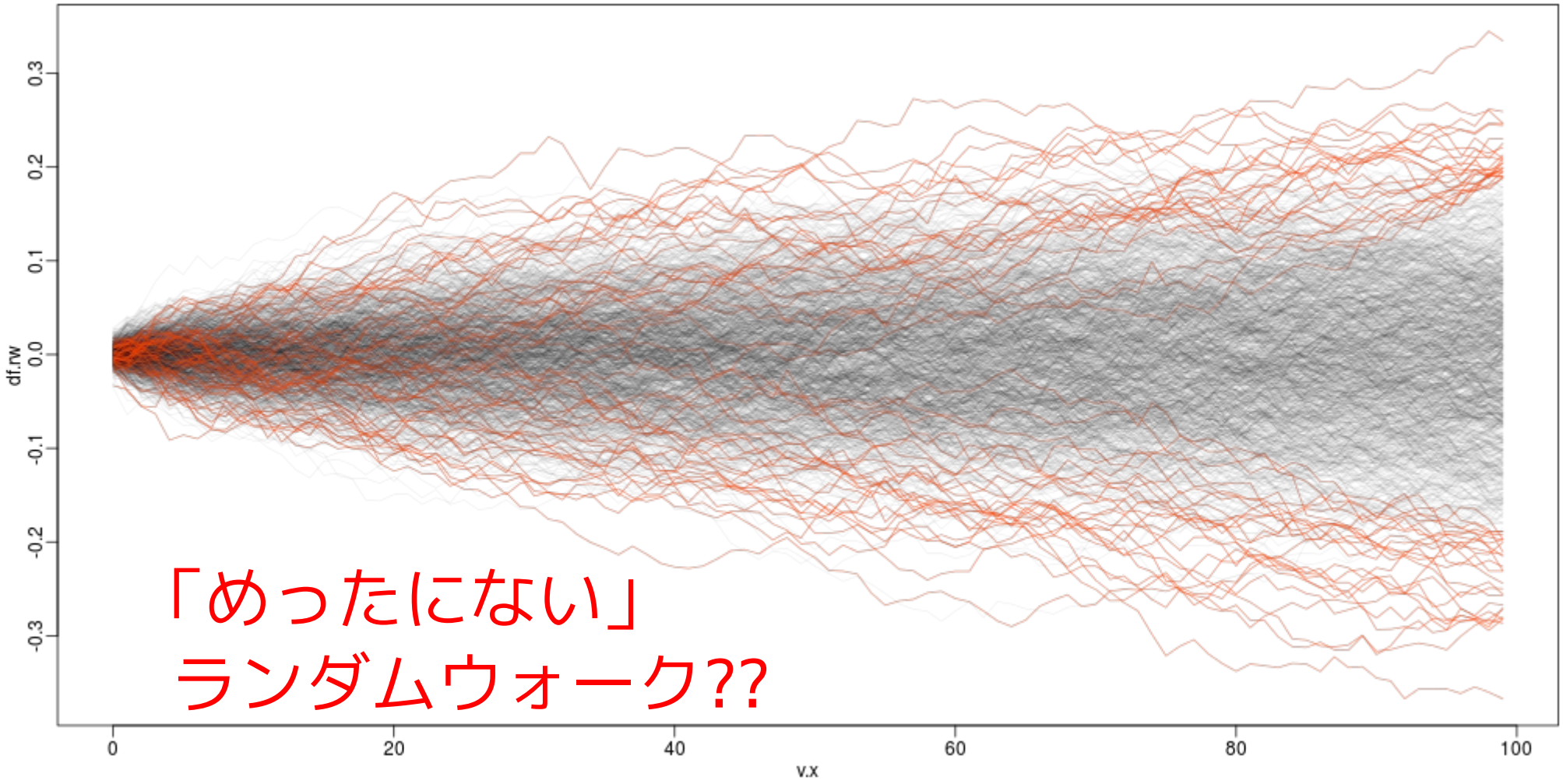
ランダムウォークなサンプル時系列

とりあえず 1000 本ほど生成してみました



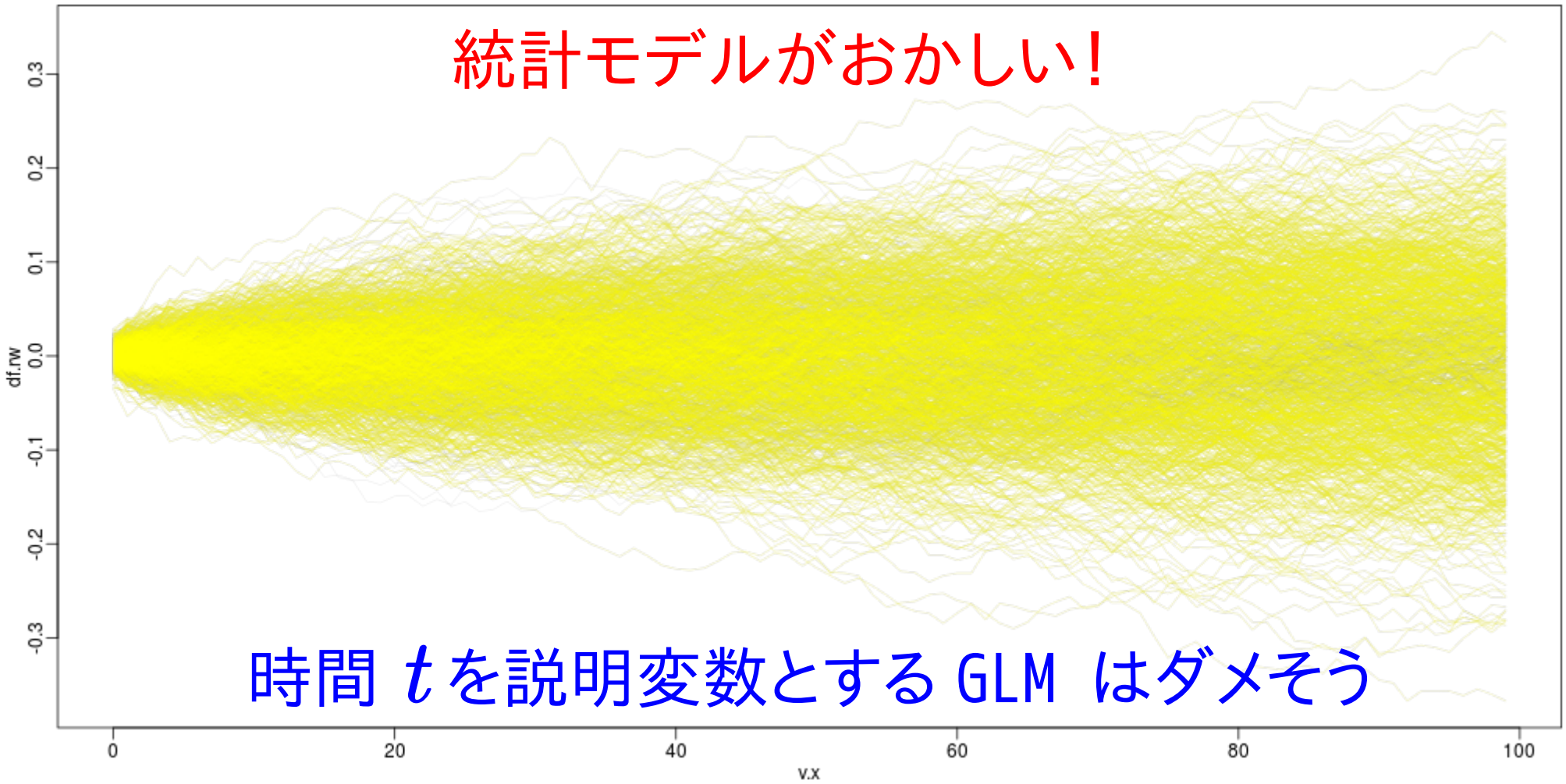
例外的な時系列というのはありえる

たとえば $t = 100$ でかなり外れている 50 本



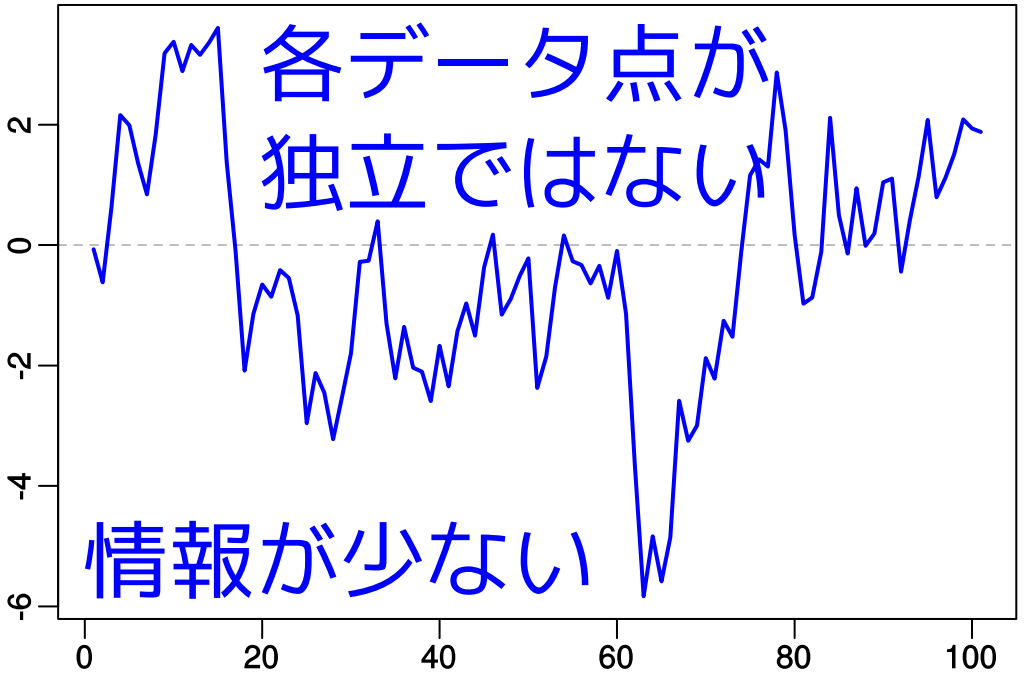
しかし直線回帰 GLM あてはめると…

ほとんどすべての場合で「ゆーい」！

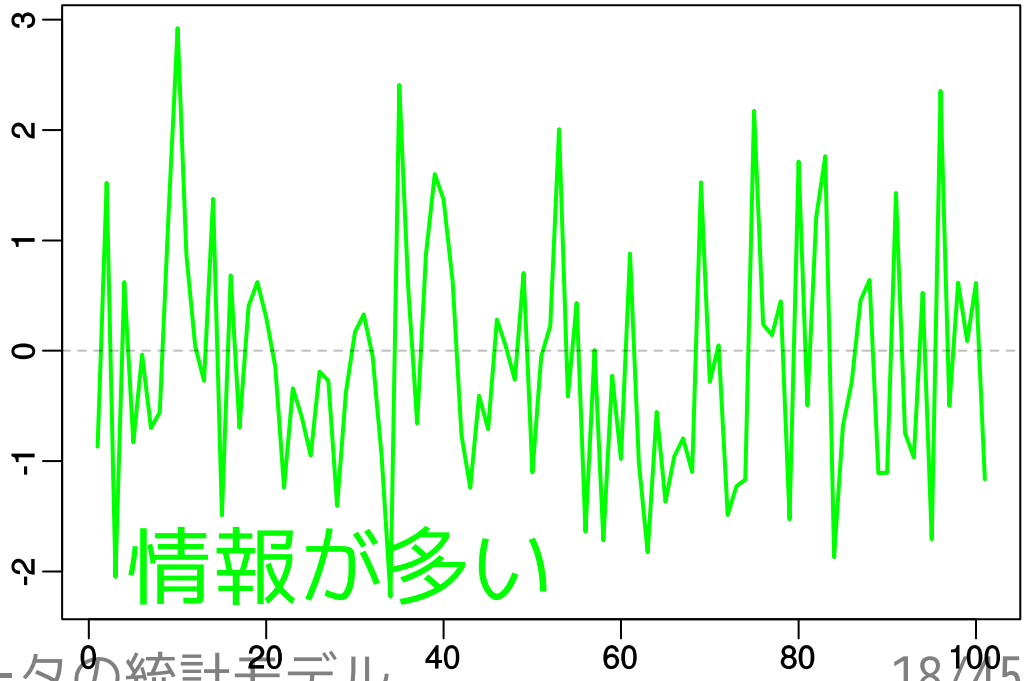


ちょっとでも傾いてたら「ゆーい」

実際には
こんなデータ
なのに



R の glm() は
こんなデータ
だとみなしている



時間的自己相関

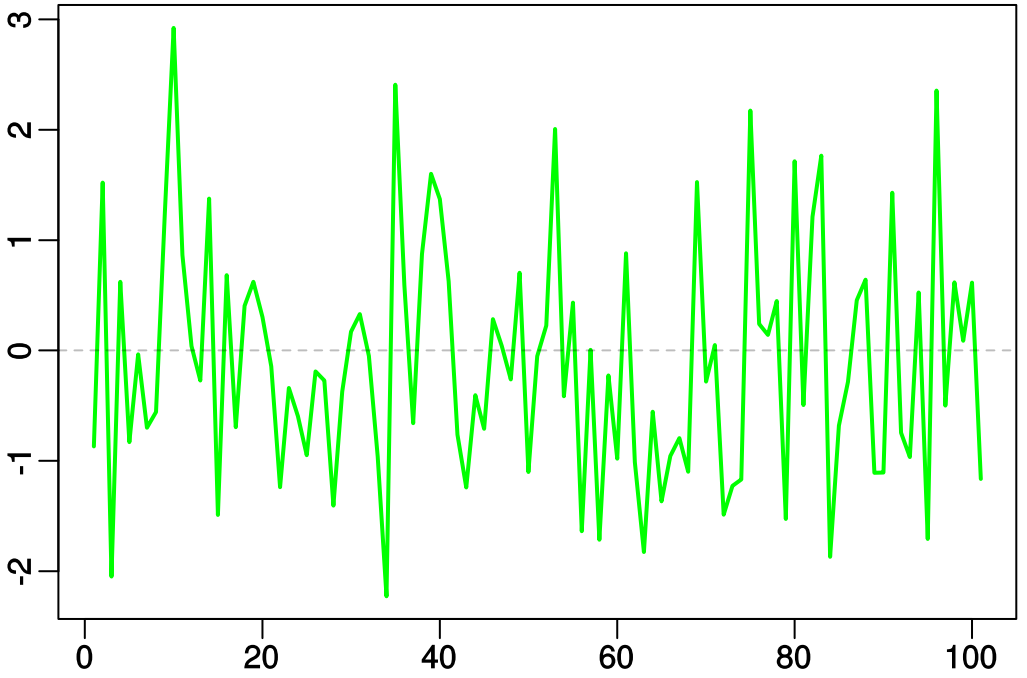
（略称:自己相関，時間相関）

を調べたらいいの?

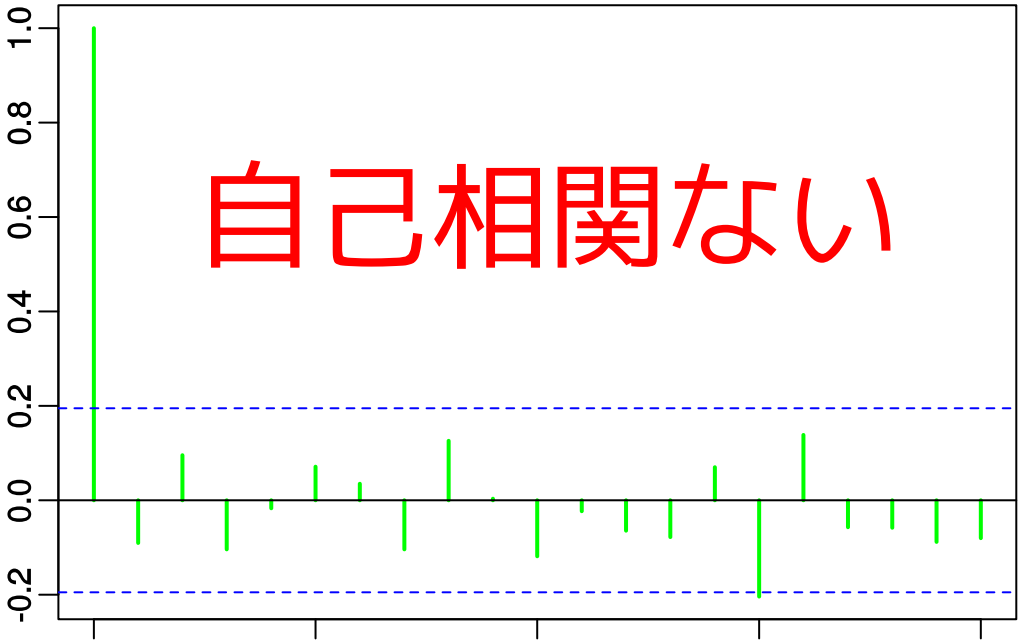
$$\rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-1})}}$$

R の ts クラス: 時系列をあつかう

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

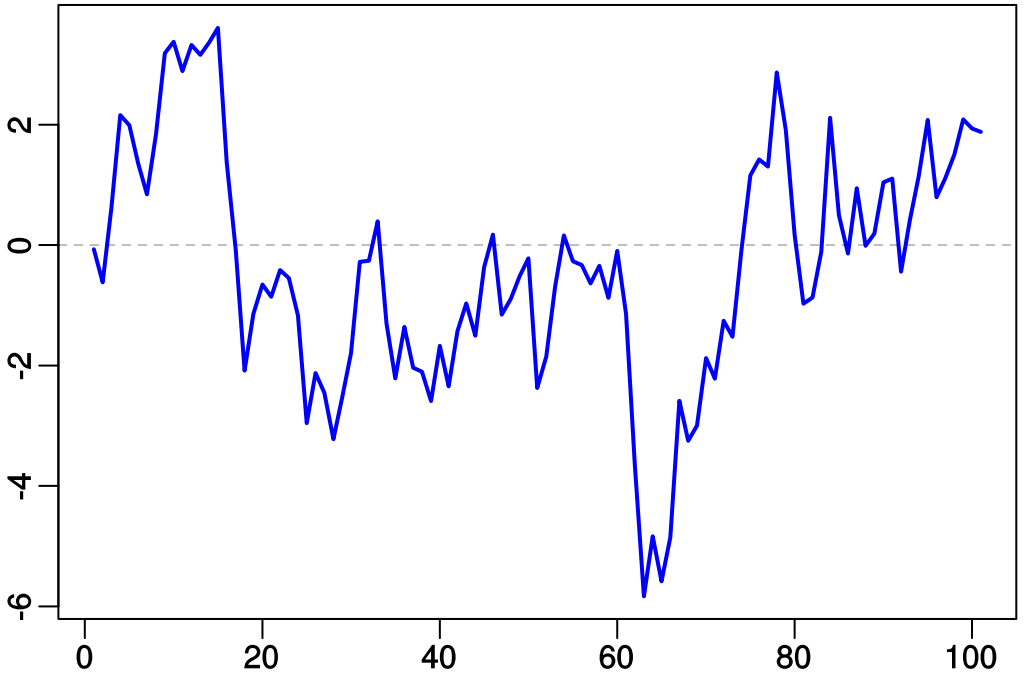



```
plot(acf(ts(Y)))
```

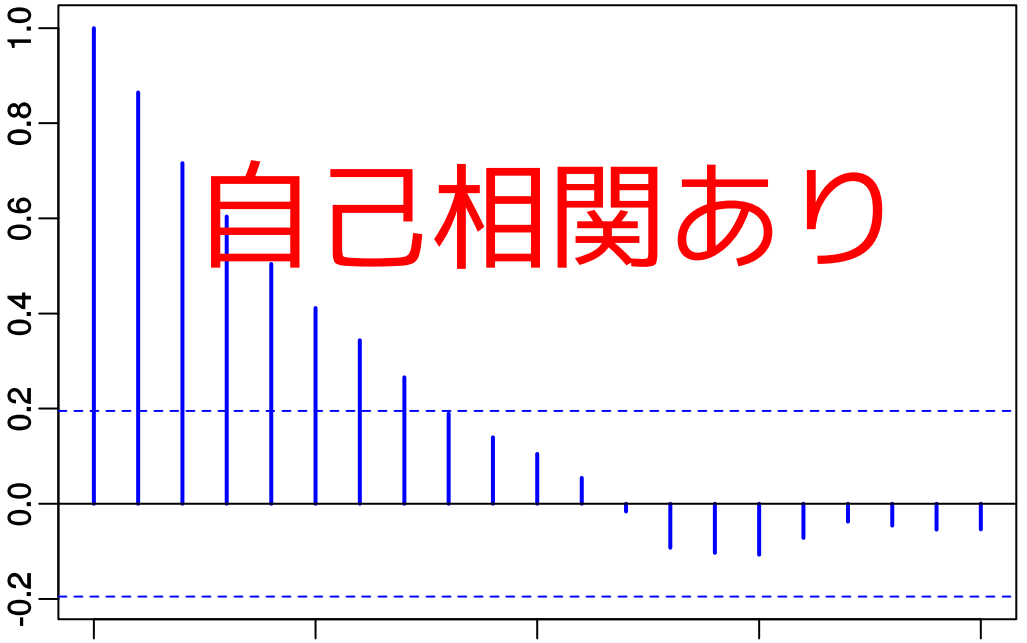


自己相関減衰の様子を図示

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

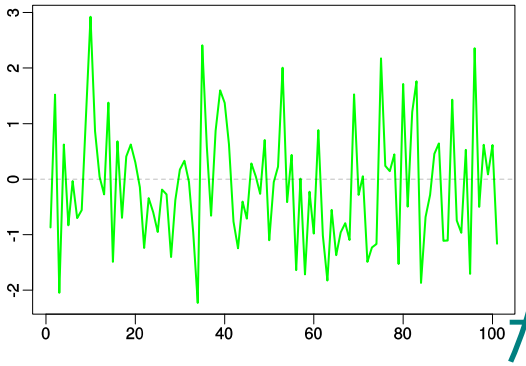



```
plot(acf(ts(Y)))
```

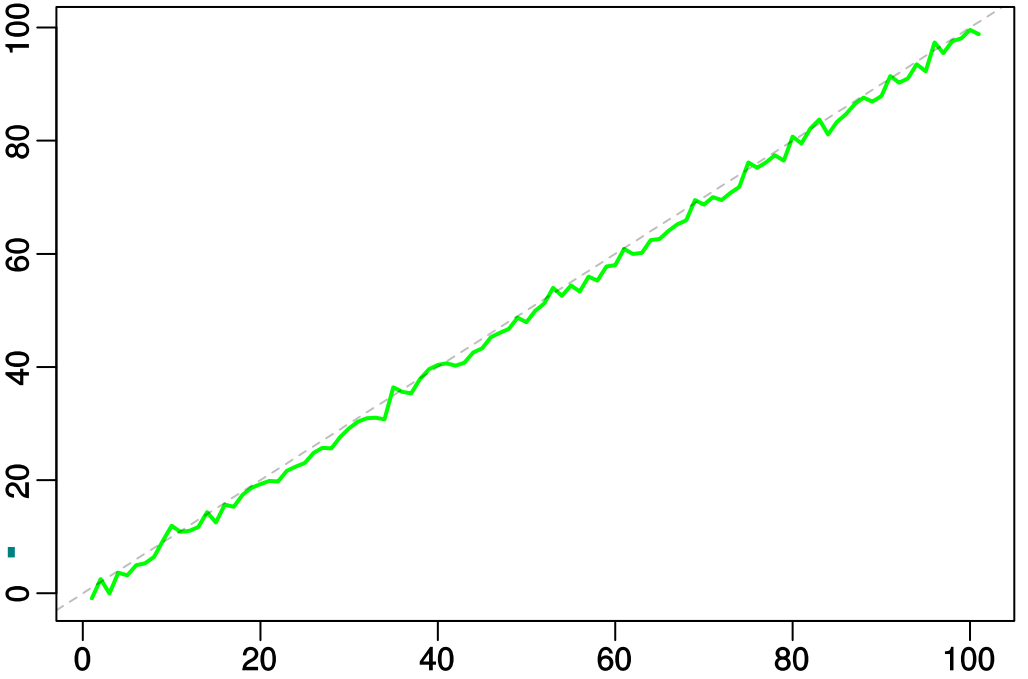


各点独立のデータをナナメにすると?

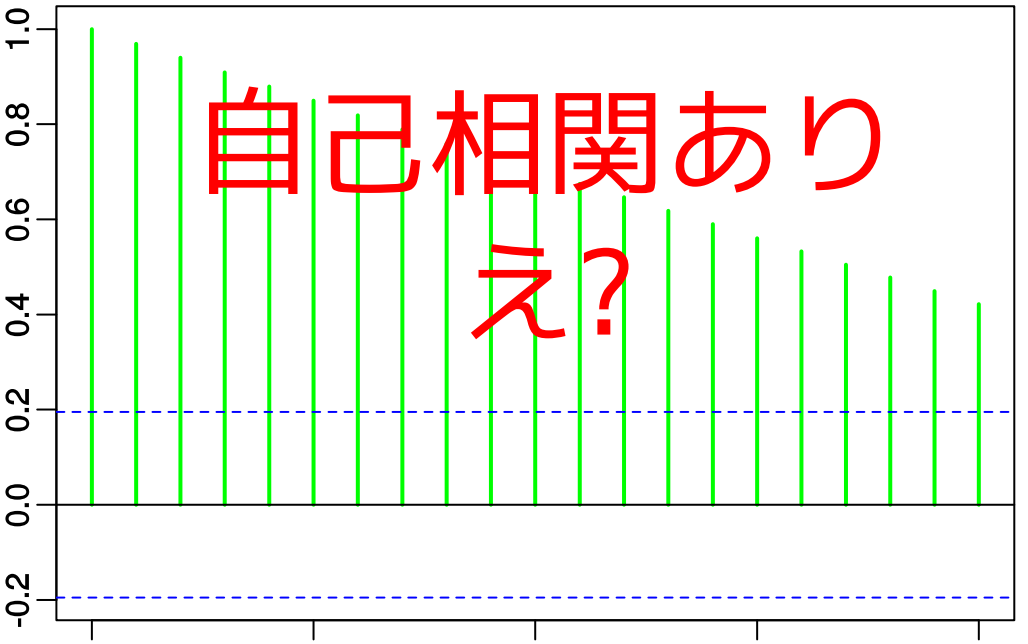
```
plot(ts(Y))
```



これを
ナナメに
したもの
なんだけど…



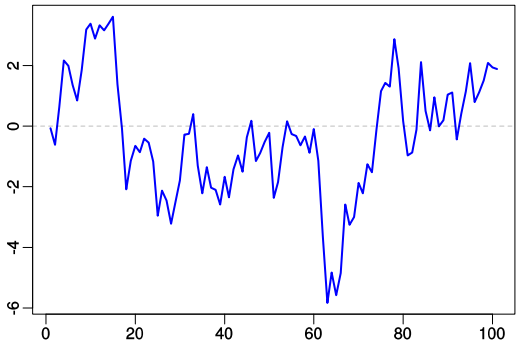

```
plot(acf(ts(Y)))
```



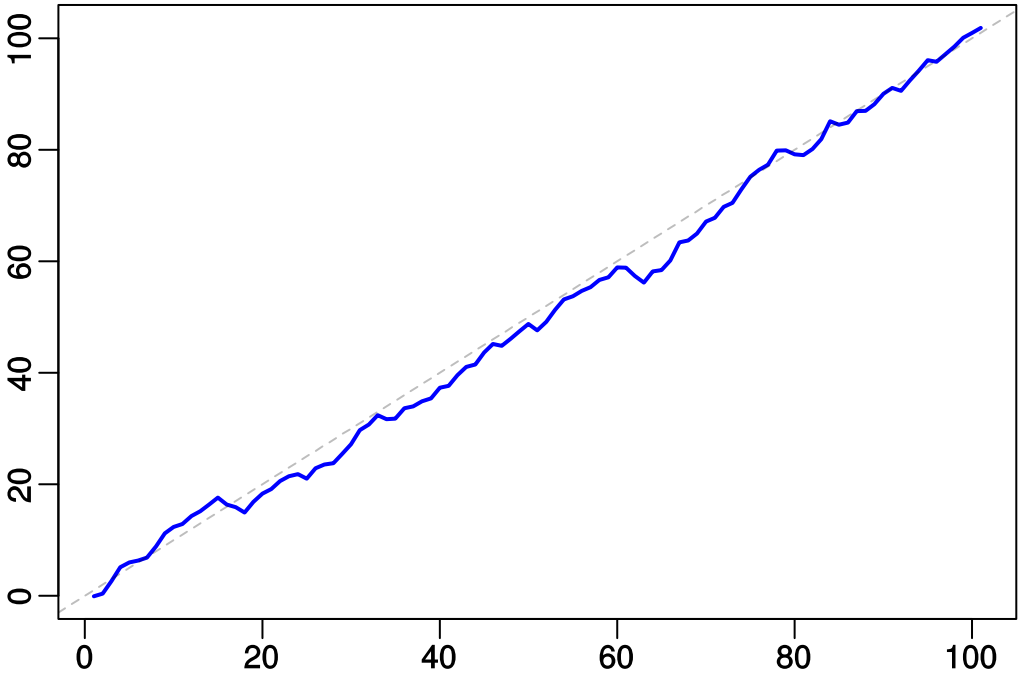
自己相関あり
え?

各点独立のデータをナナメにすると?

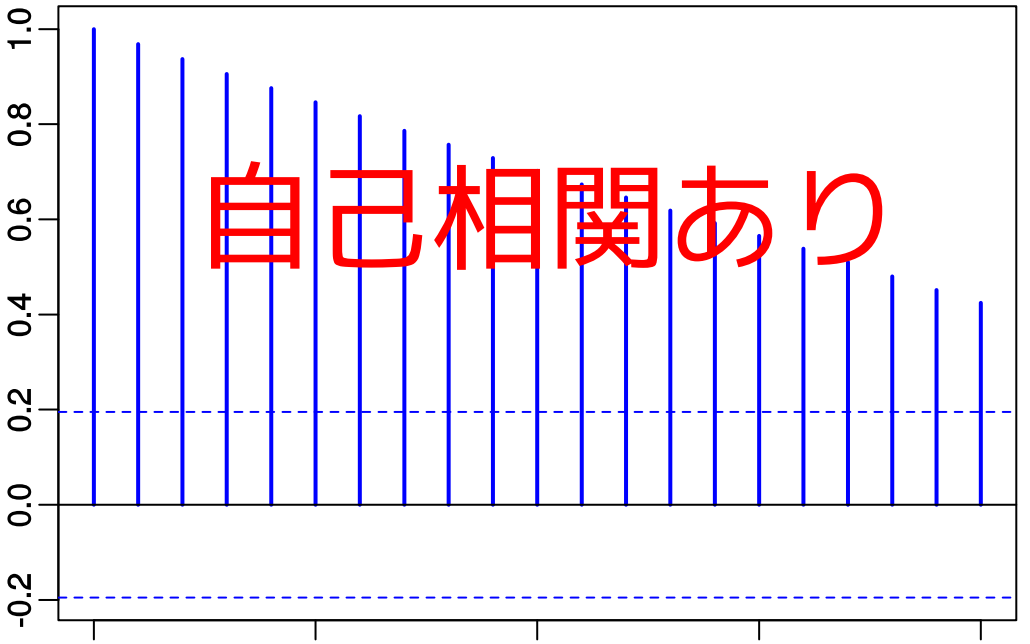
```
plot(ts(Y))
```



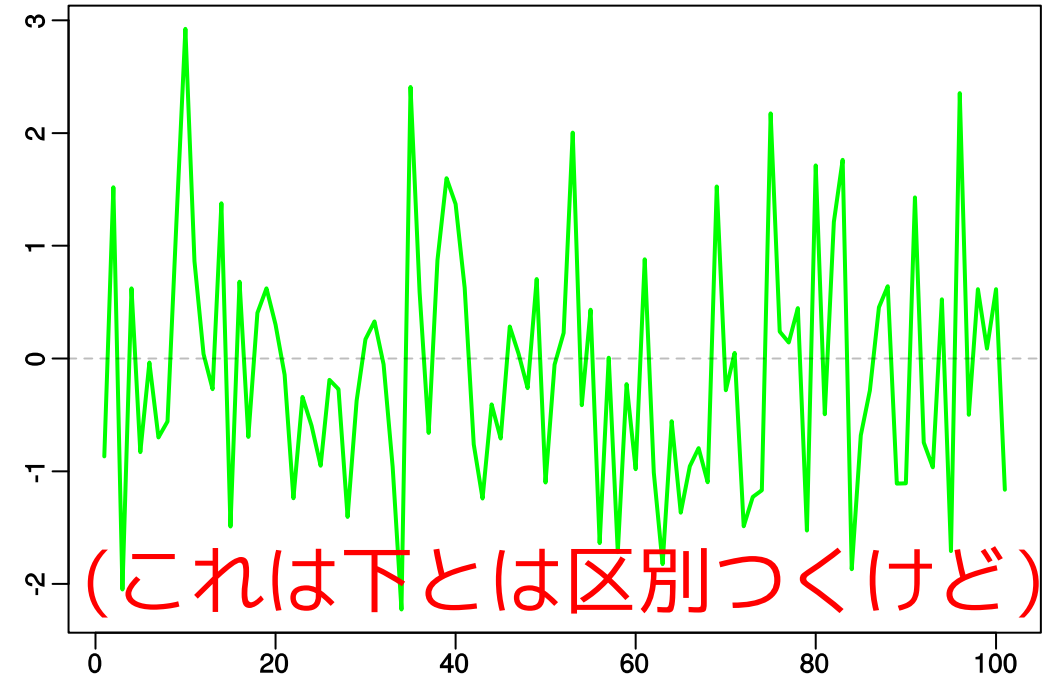
これを
ナナメに
したもの



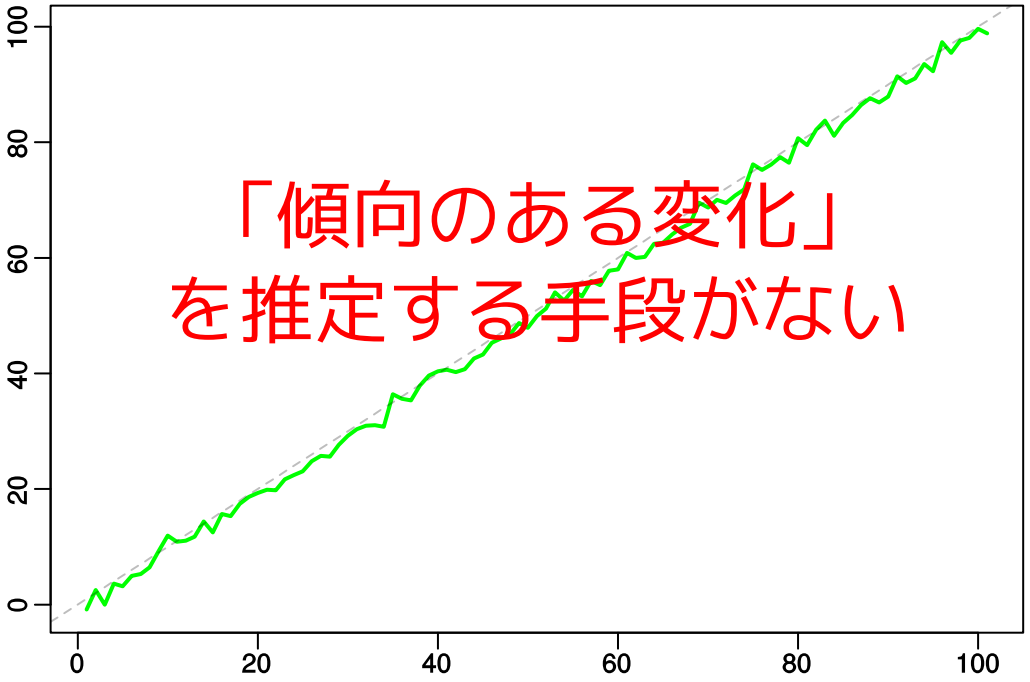

```
plot(acf(ts(Y)))
```



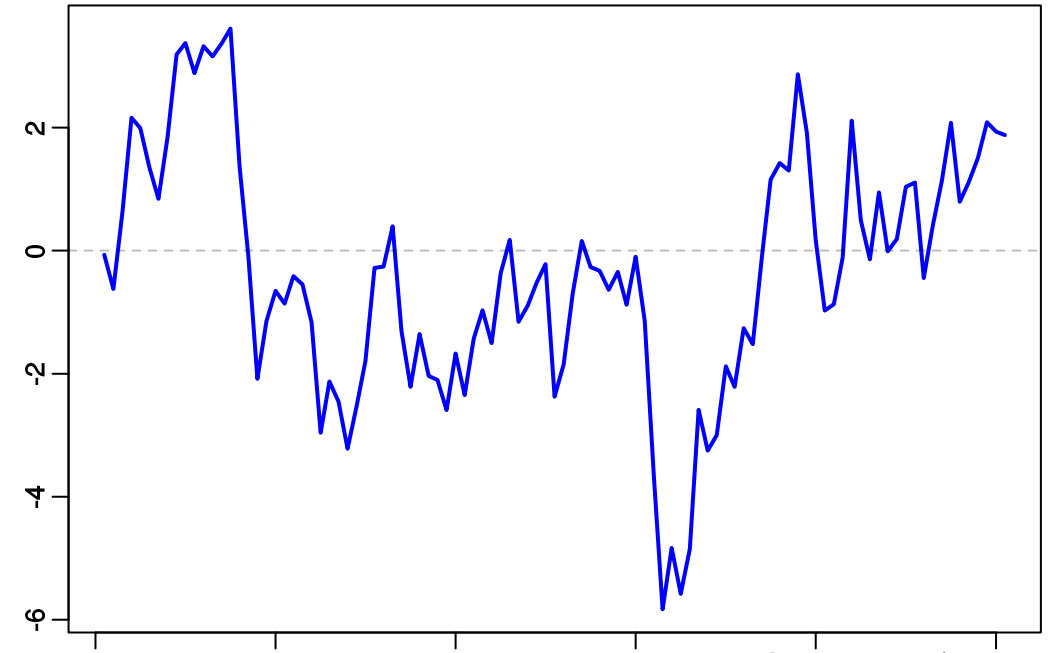
自己相関係数みても区別がつかない

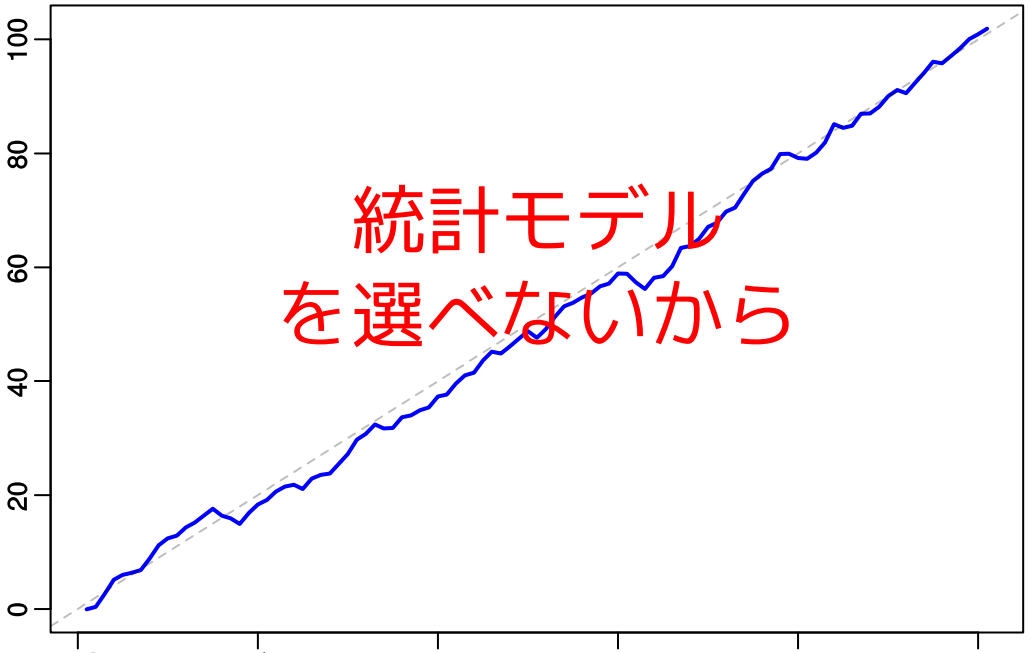


(これは下とは区別つくけど)



「傾向のある変化」
を推定する手段がない





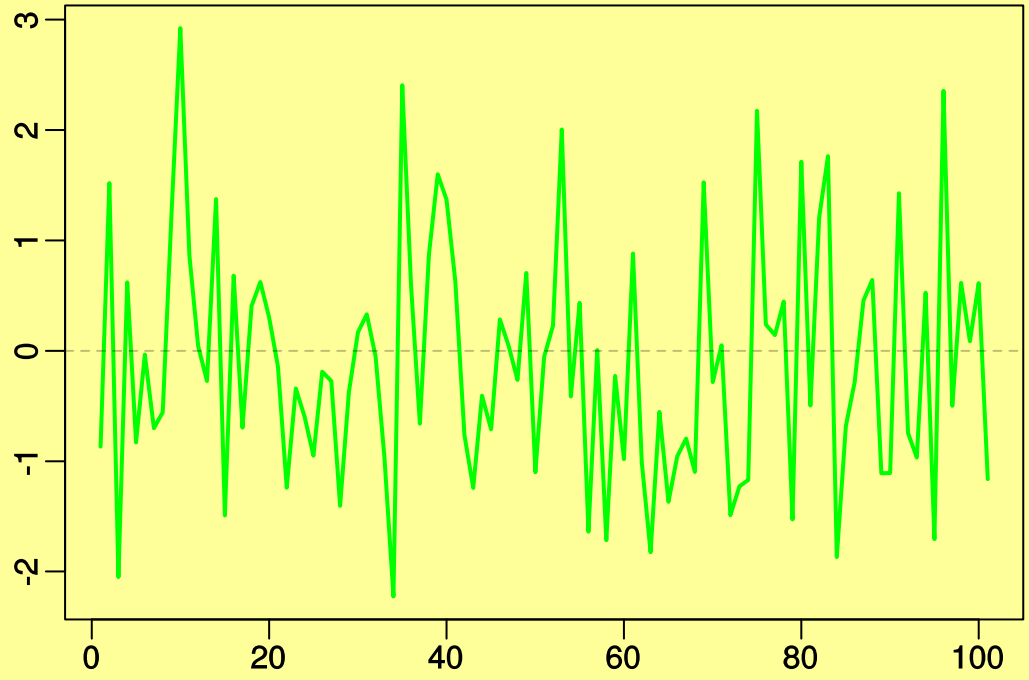
統計モデル
を選べないから

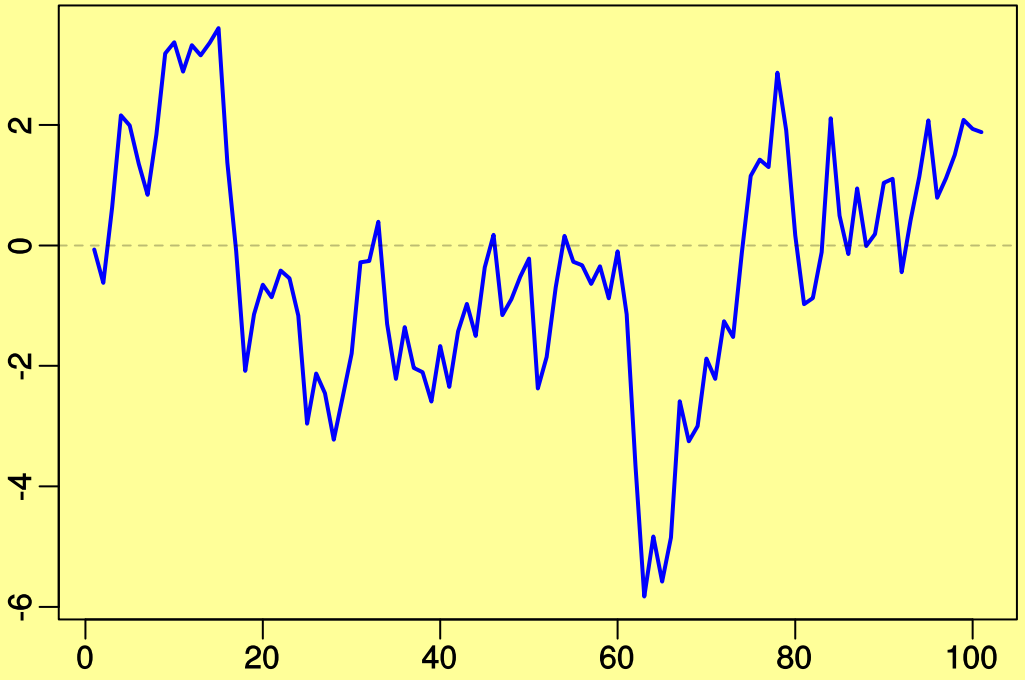
状態空間モデルでたちむかう

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?





「統計モデル」とは何か?

どんな統計解析においても
統計モデルが使用されている

- 観察によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり，これはデータにみられるばらつきを表現する手段である
- データとモデルを対応づける手つづきが準備されていて，モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる

「統計モデル」のしくみを理解しよう!

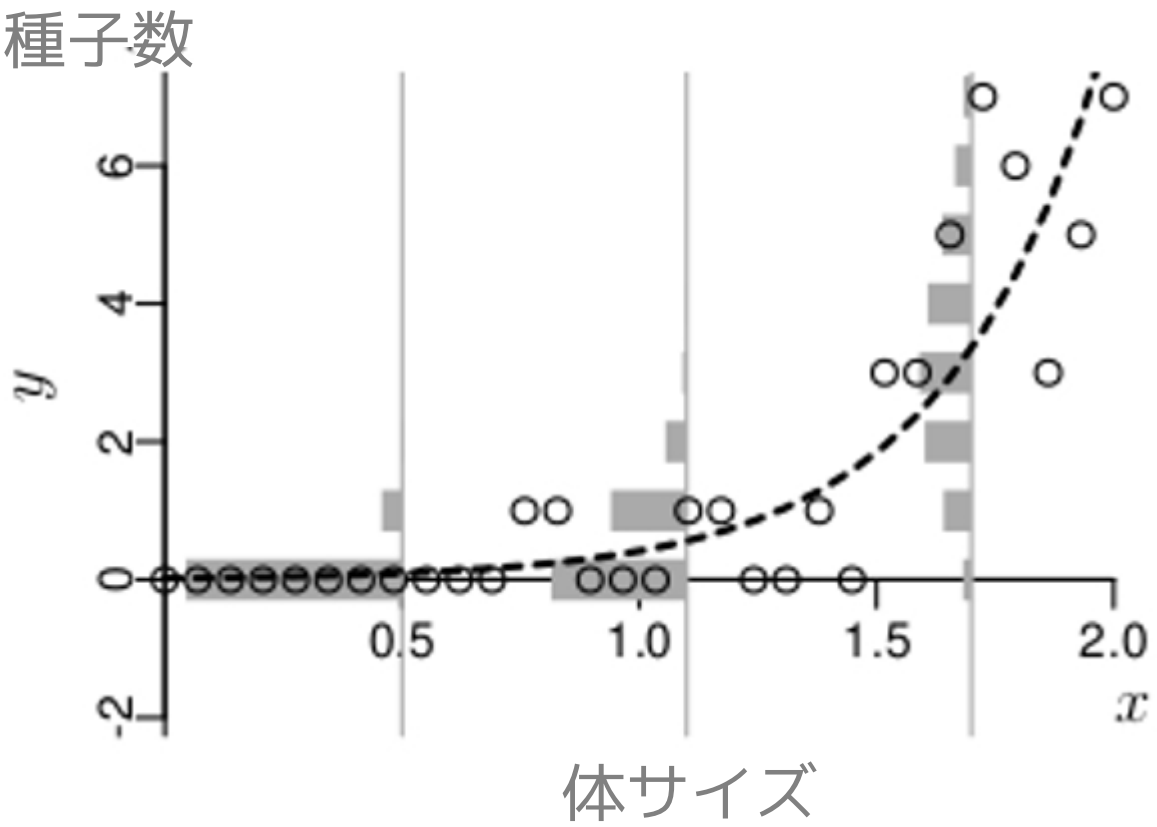
もうすこし「わかった」ような気分?

種子数の平均値はサイズ x とともに増大する

➡ どのように変化するのか?
数式で書くとどうなる?

平均値が増大するとばらつきが変化する

➡ どのようにばらつくのか?
確率分布?

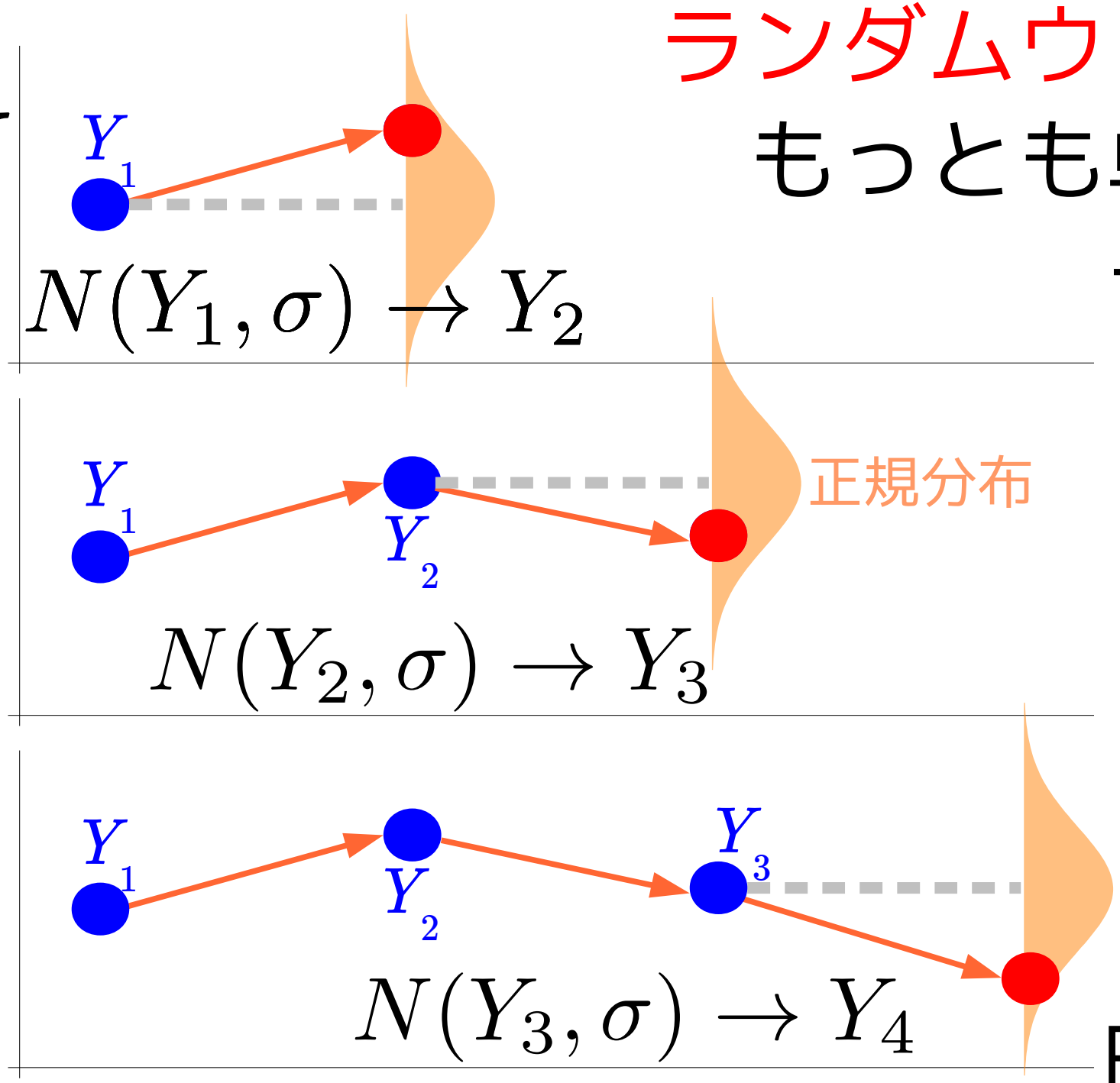


統計モデルをデータにうまくあてはめる

➡ どのようにあてはめるのが妥当なのか? パラメーター推定法?

ランダムウォーク

もっとも単純なモデル

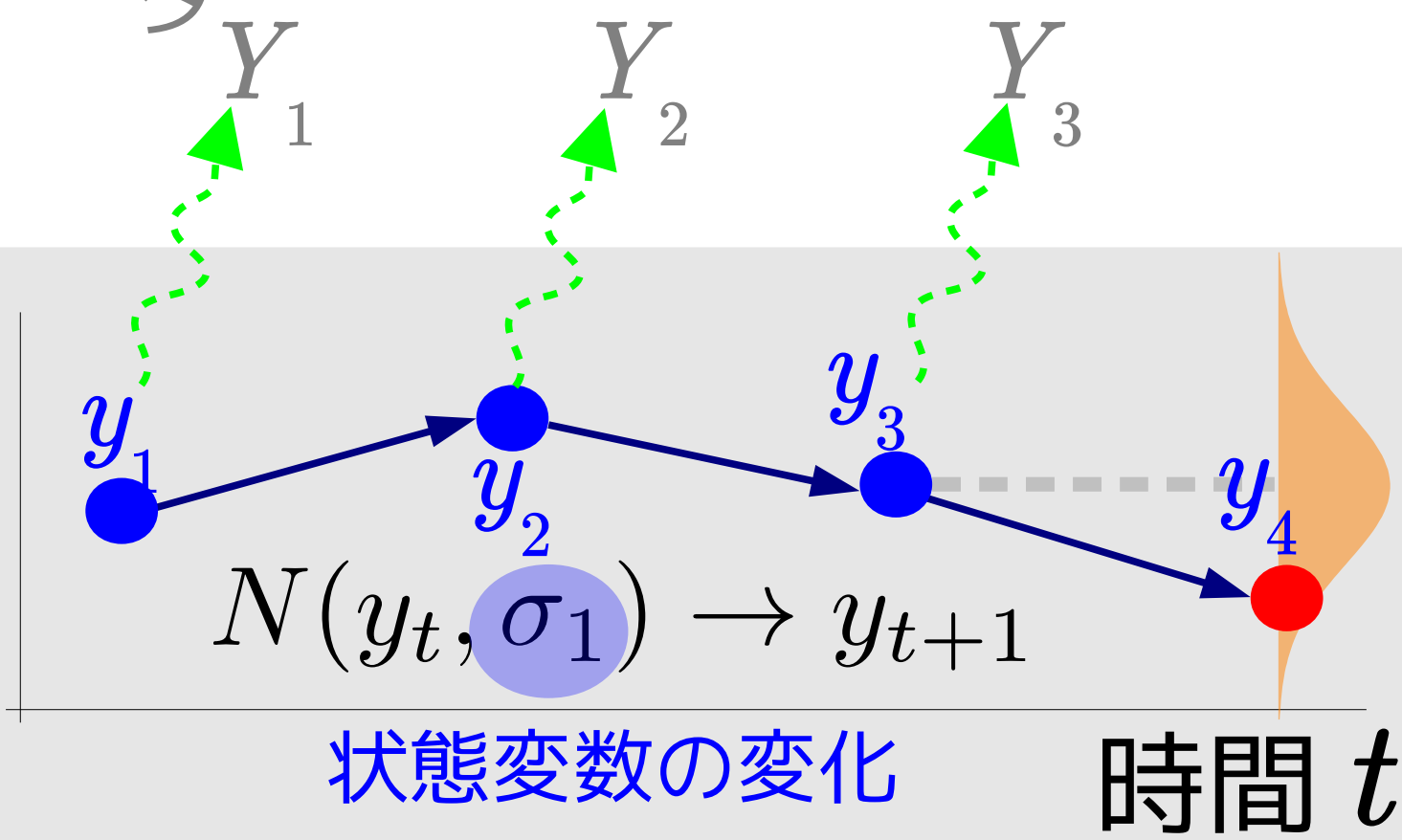


状態空間モデル

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



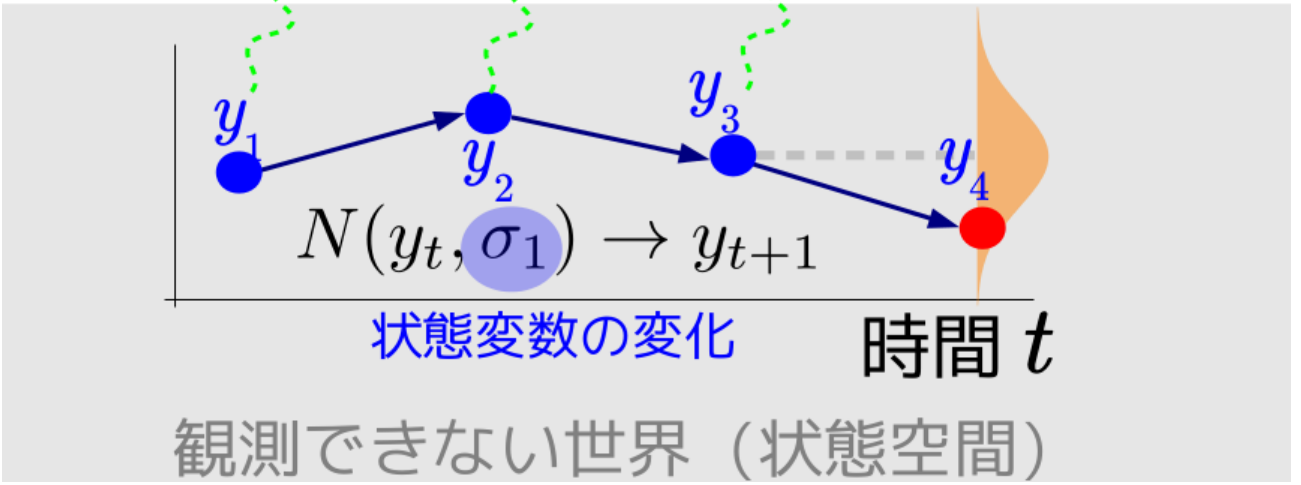
観測できない世界（状態空間）

状態空間モデル

観測の誤差

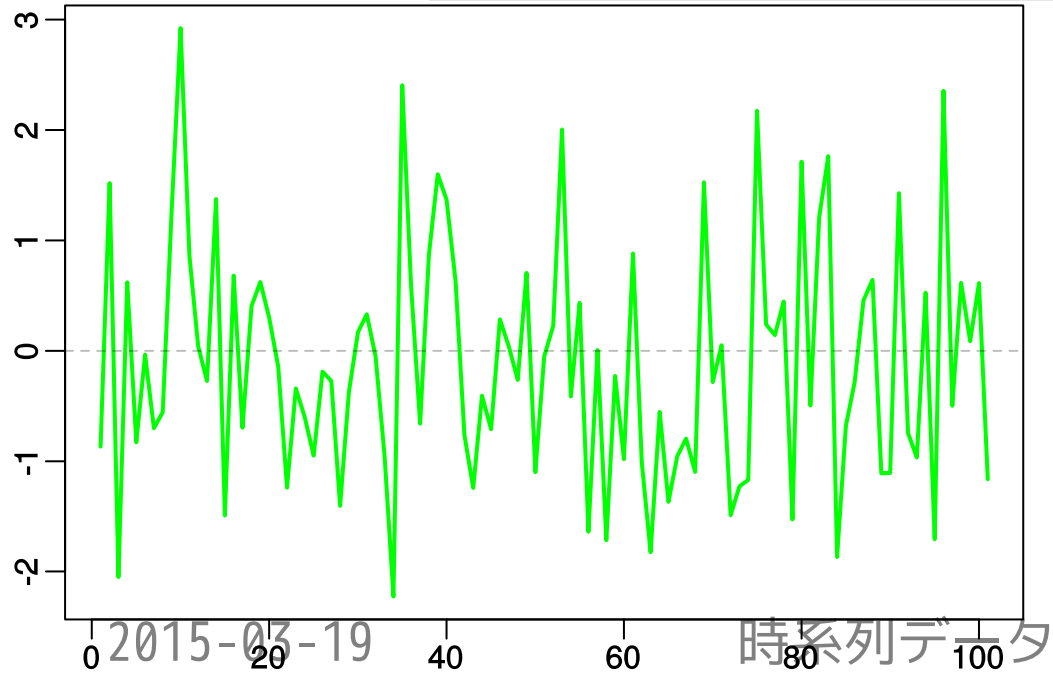
$N(y_t, \sigma_2) \to Y_t$ 二種類の σ をもつ

観測データ Y_1 Y_2 Y_3

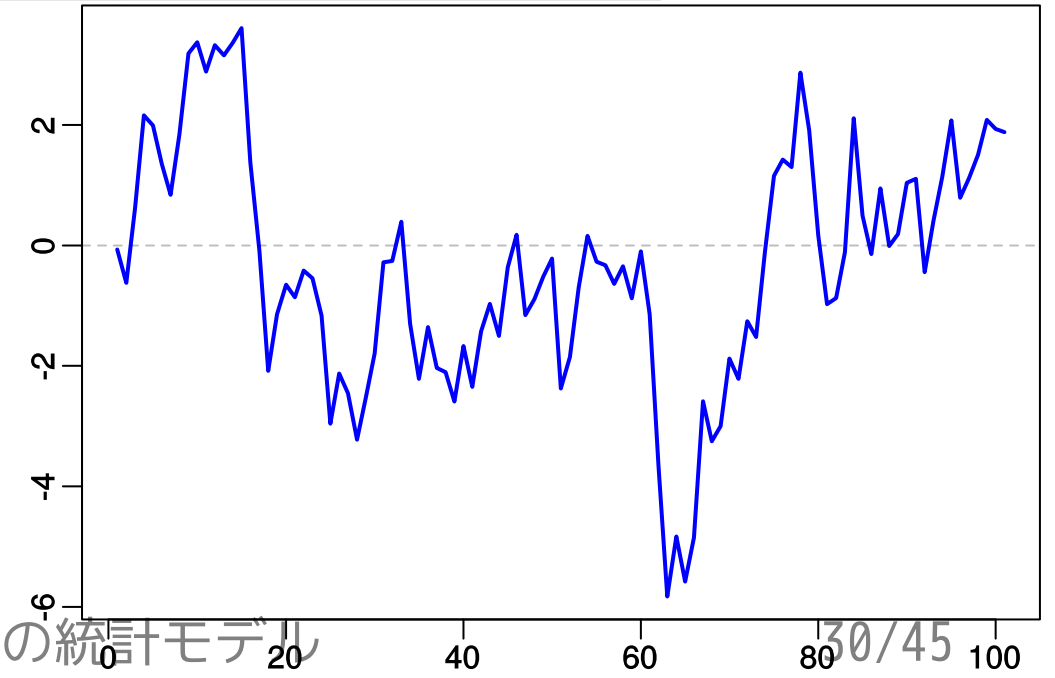


観測できない世界（状態空間）

σ_2 大
σ_1 小



σ_2 小
σ_1 大

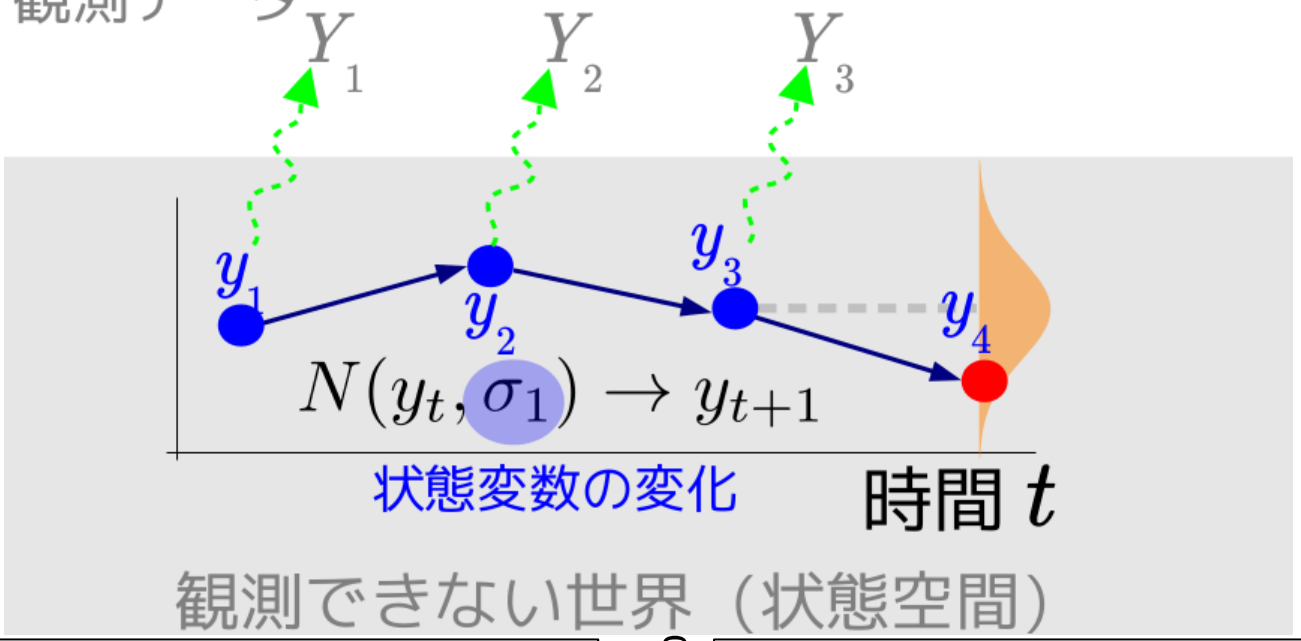


状態空間モデル

観測の誤差

$N(y_t, \sigma_2) \to Y_t$ 二種類の σ をもつ

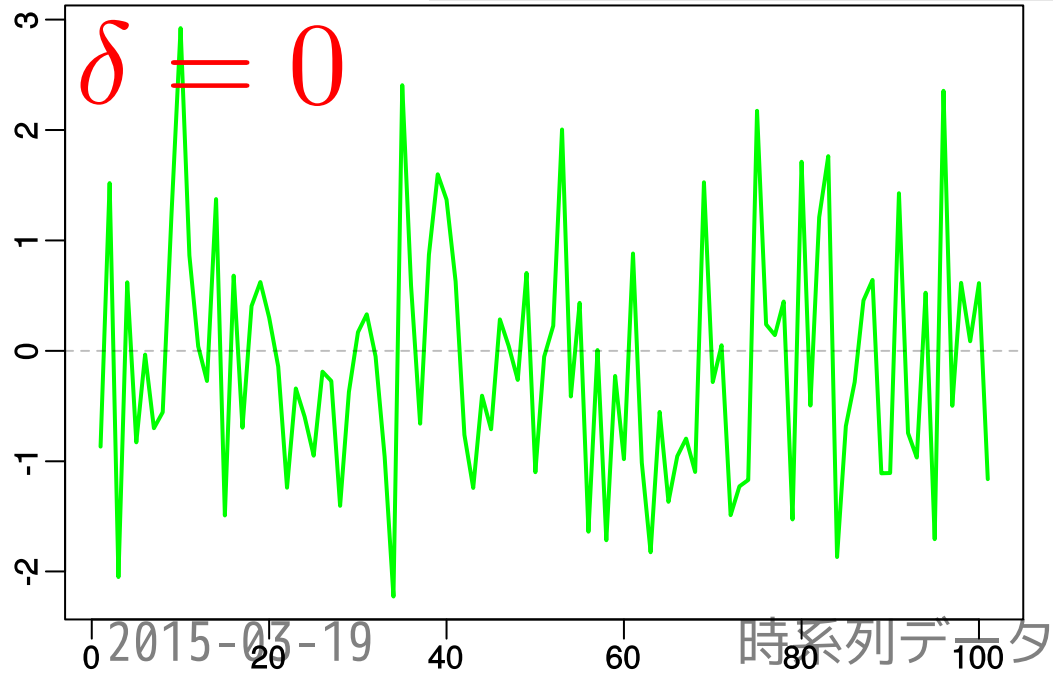
観測データ

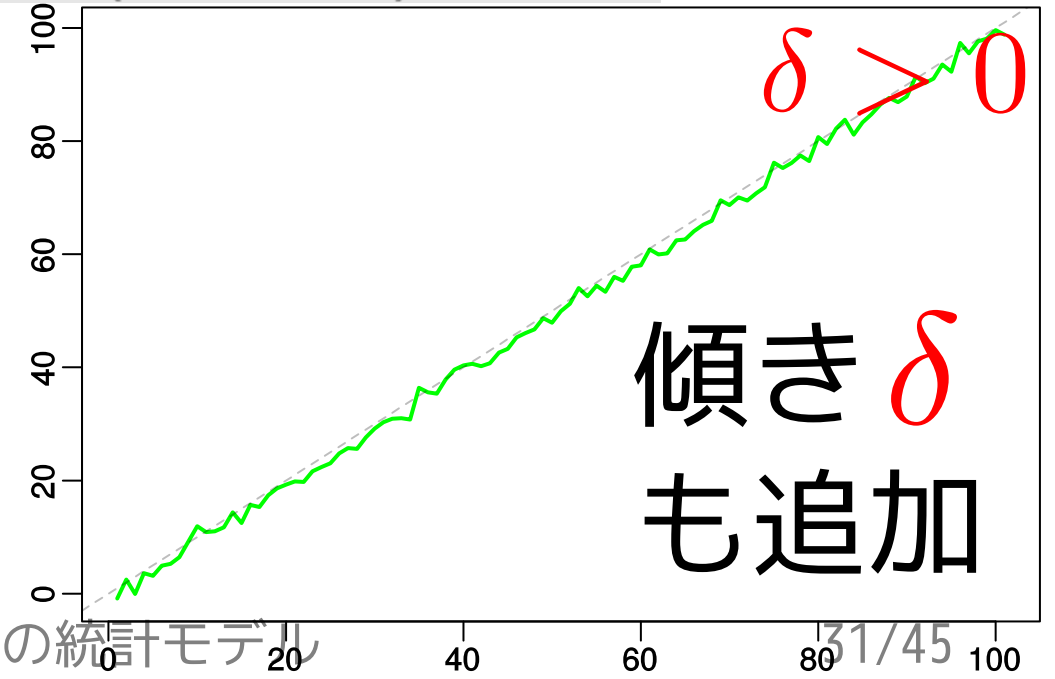


観測できない世界（状態空間）

σ_2 大

σ_1 小

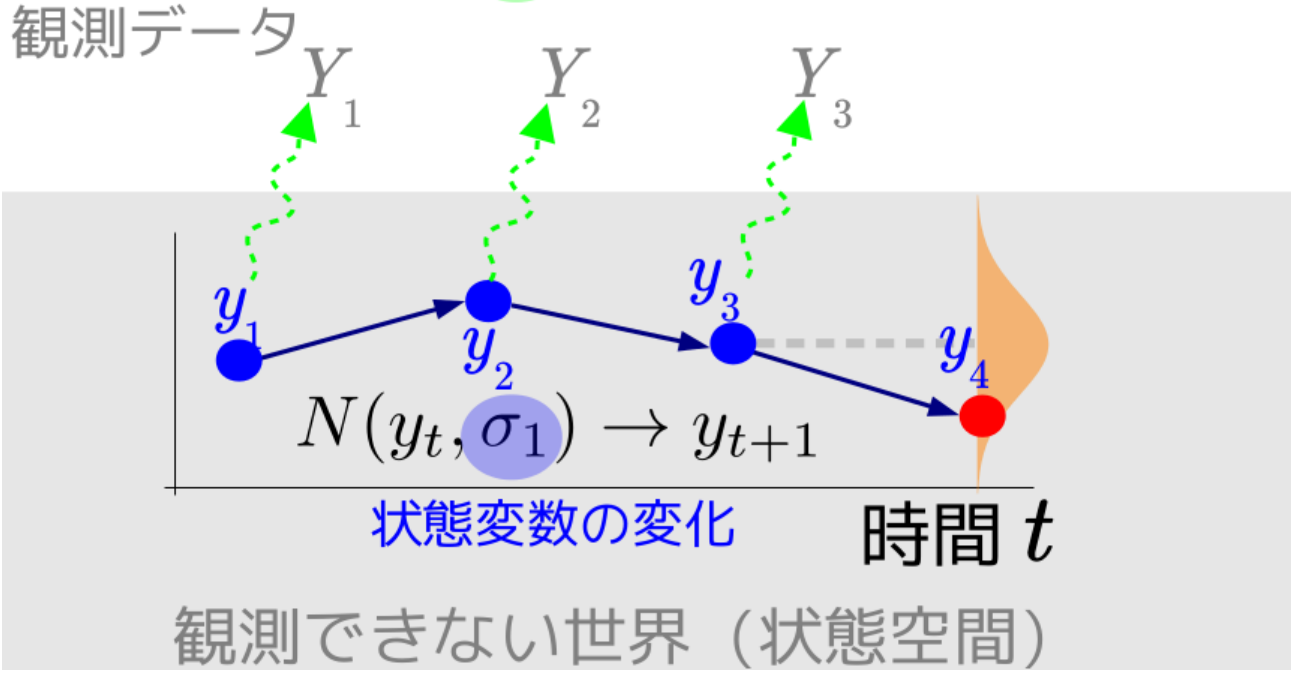




状態空間モデル

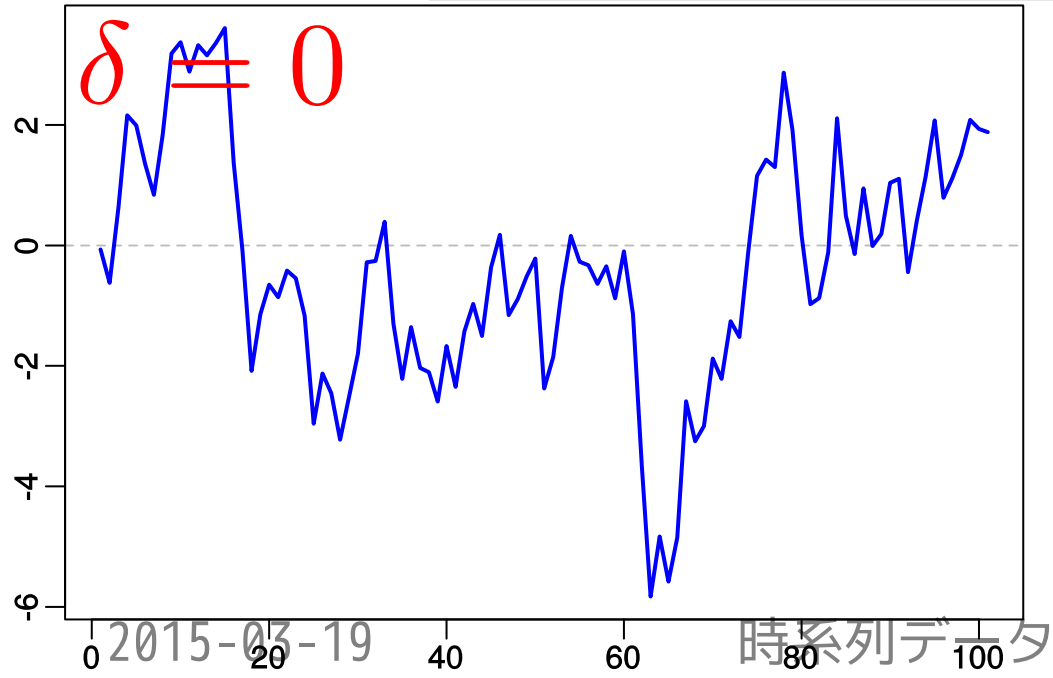
観測の誤差

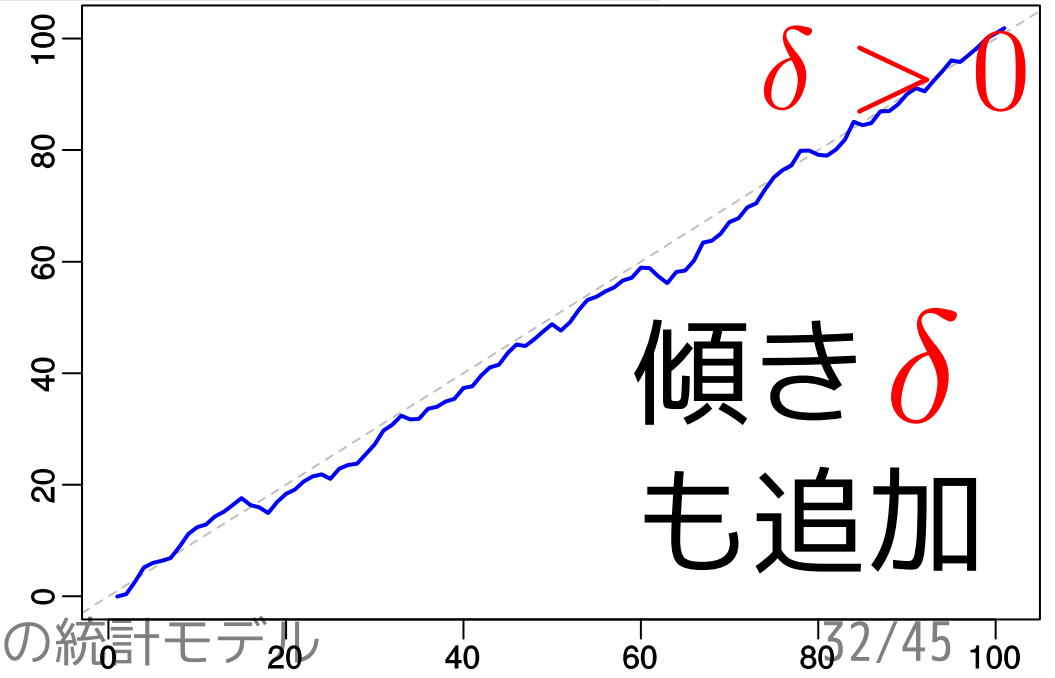
$N(y_t, \sigma_2) \to Y_t$ 二種類の σ をもつ



σ_2 小

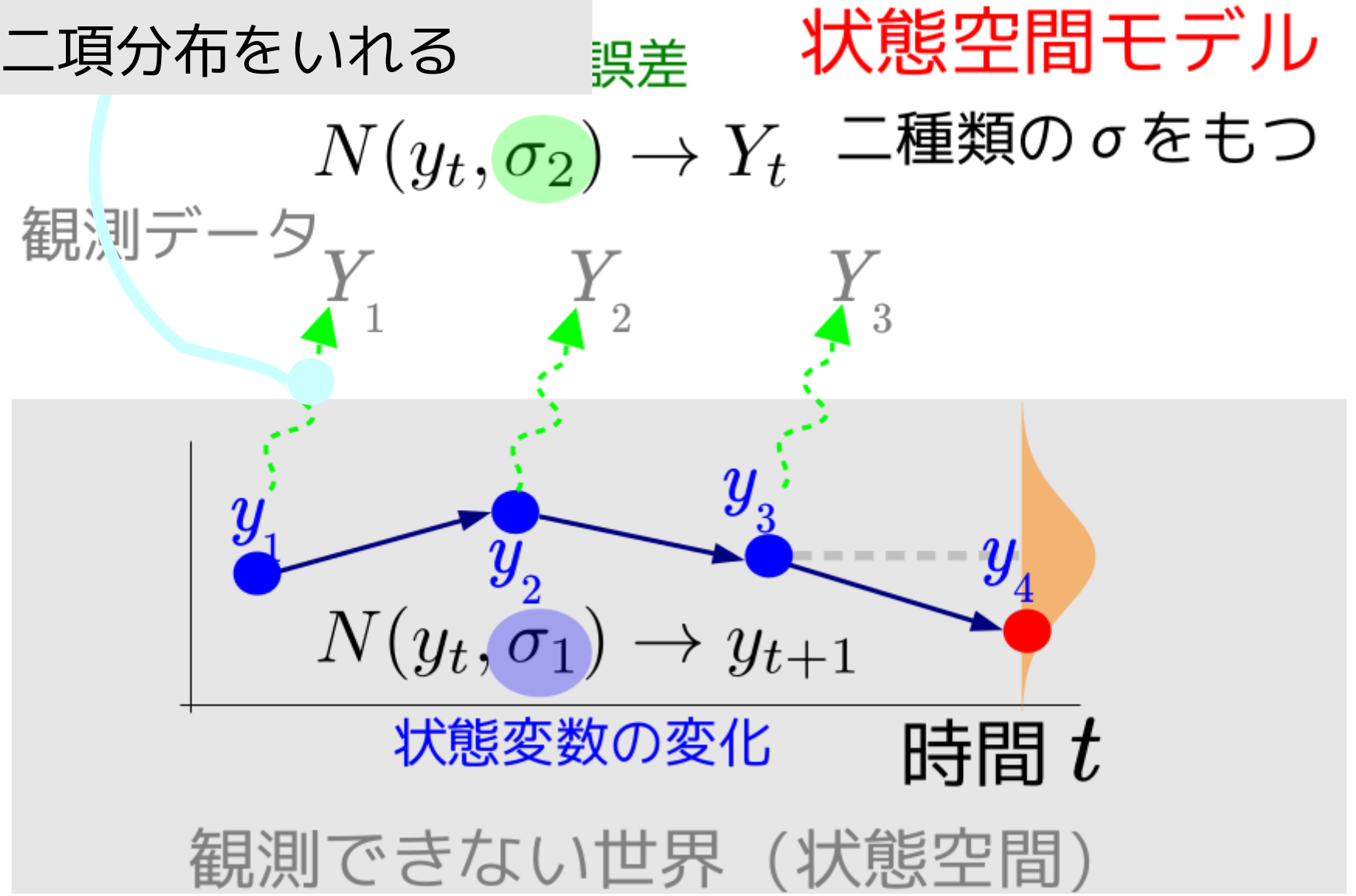
σ_1 大





状態空間モデル + GLM

この部分にポアソン分布や
二項分布をいれる



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

JAGS でいきましょう

BUGS 言語でこの単純な

モデルを記述できる

R の「したっぱ」として

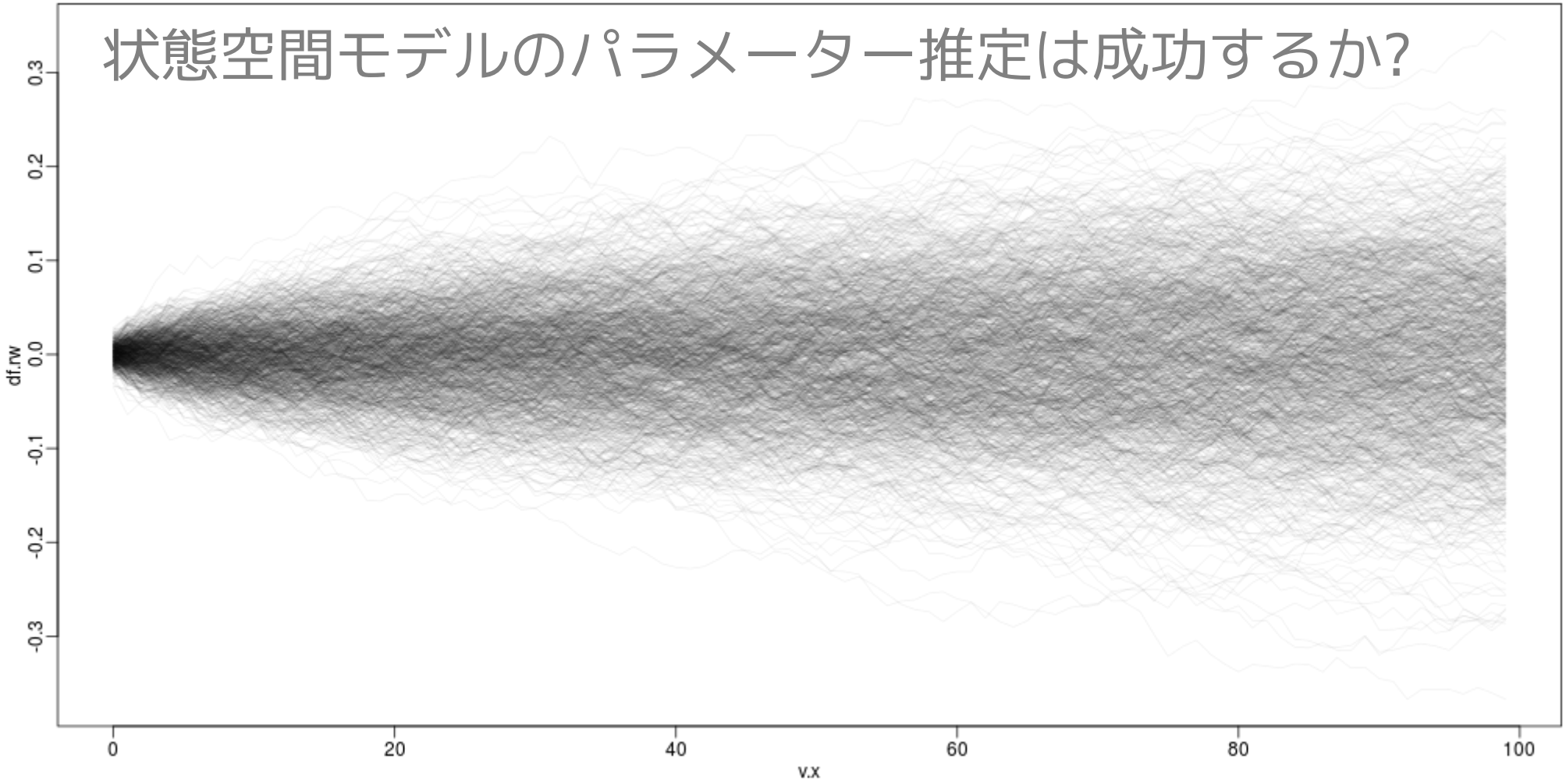
動かすことができる

時間があれば demo

```
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)
    }
}
```

1000 個の架空データを推定

いろいろなランダムウォークが生成される



状態空間モデルを

「かたむきゼロ」ランダムウォーク

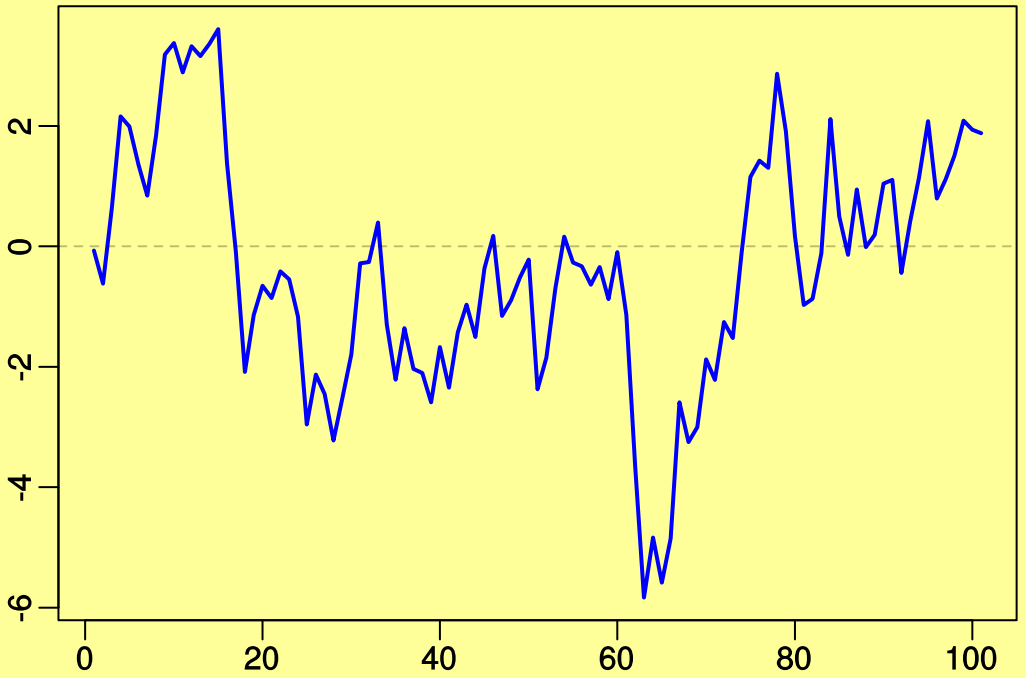
$\delta = 0$

な架空データにあてはめる

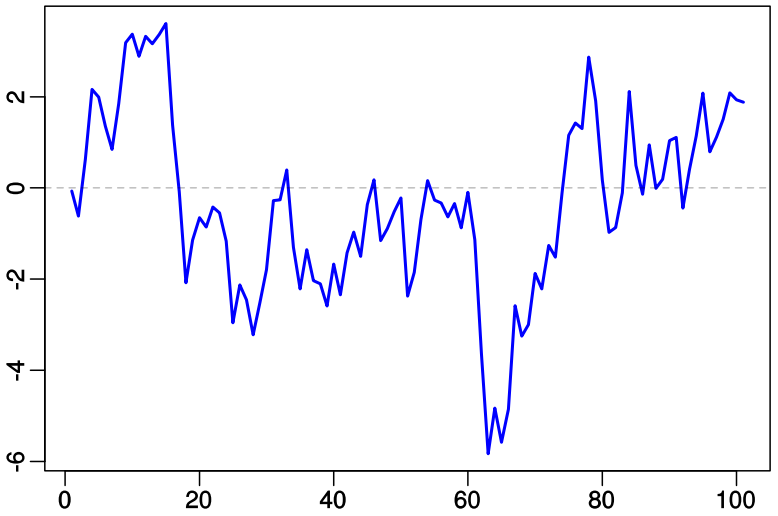
σ_2 小

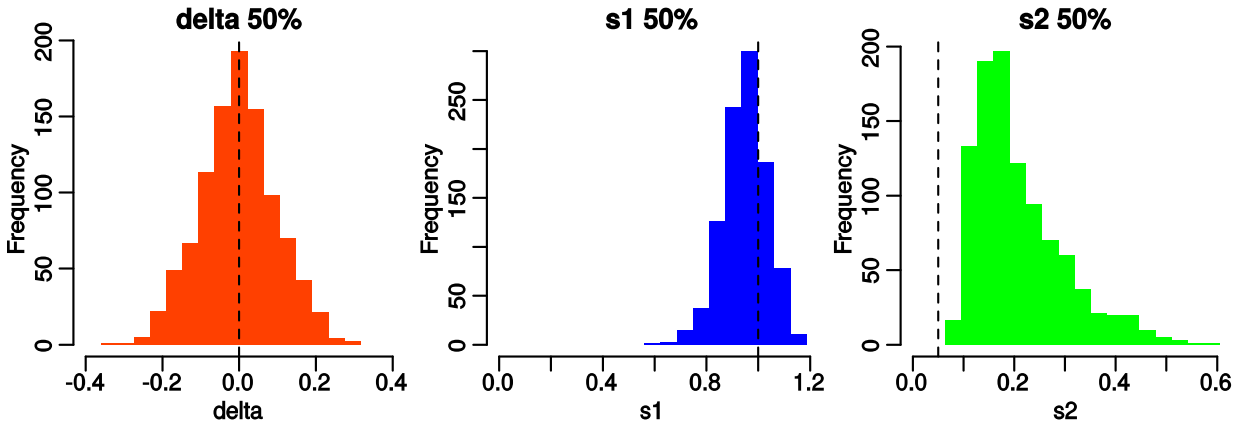
σ_1 大

$\delta = 0$

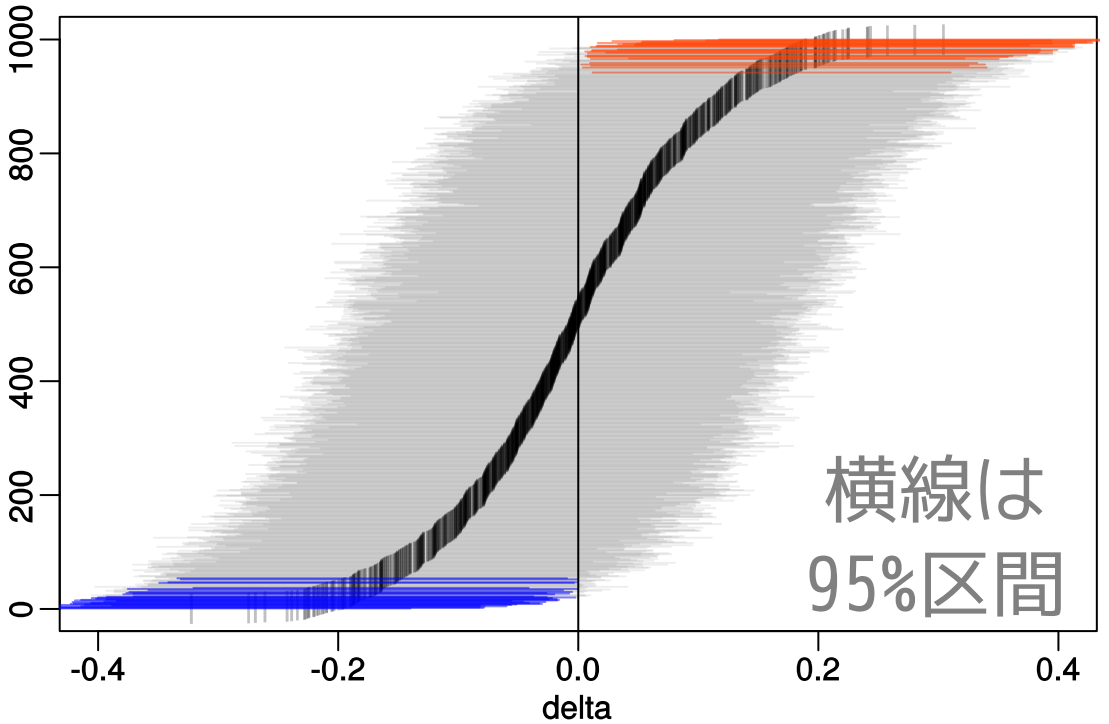


「傾き」δの事後分布を見る





真のδは 0



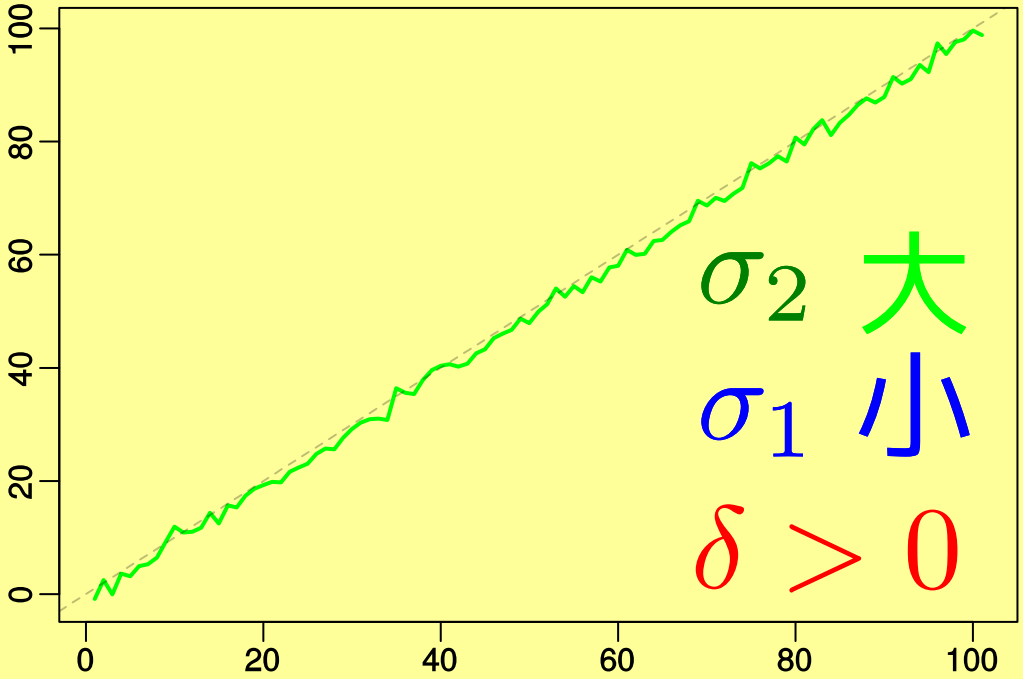
1000回中
63回ずれた

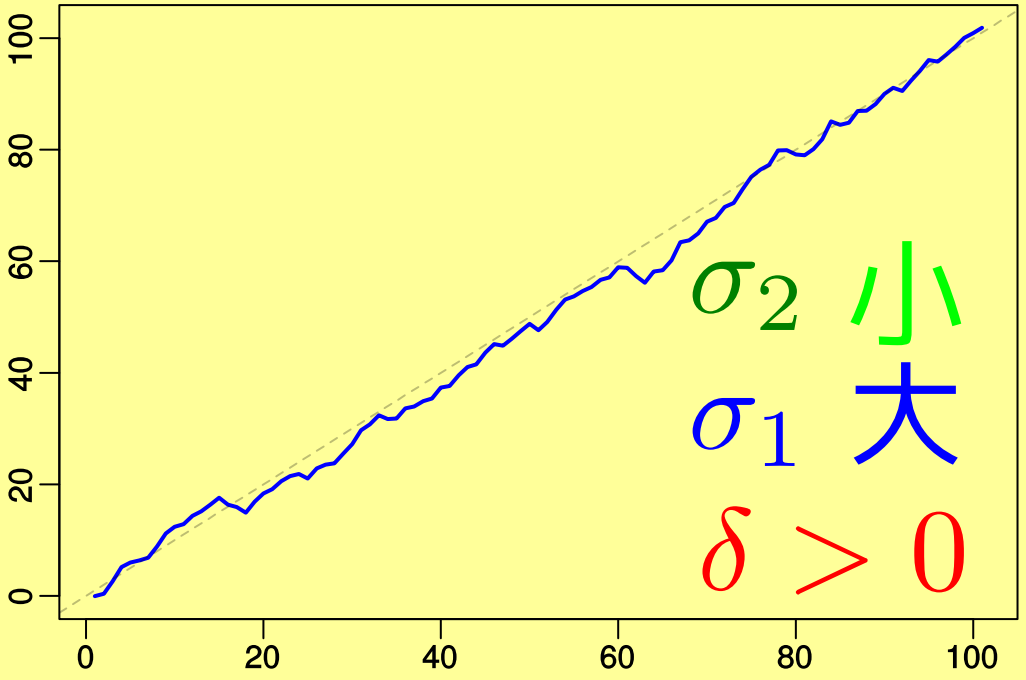
状態空間モデルを

「かたむきあり」ランダムウォーク

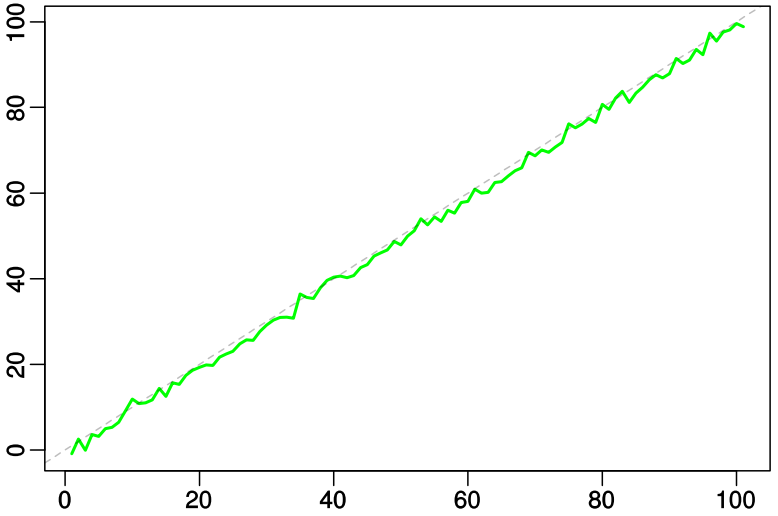
$\delta > 0$

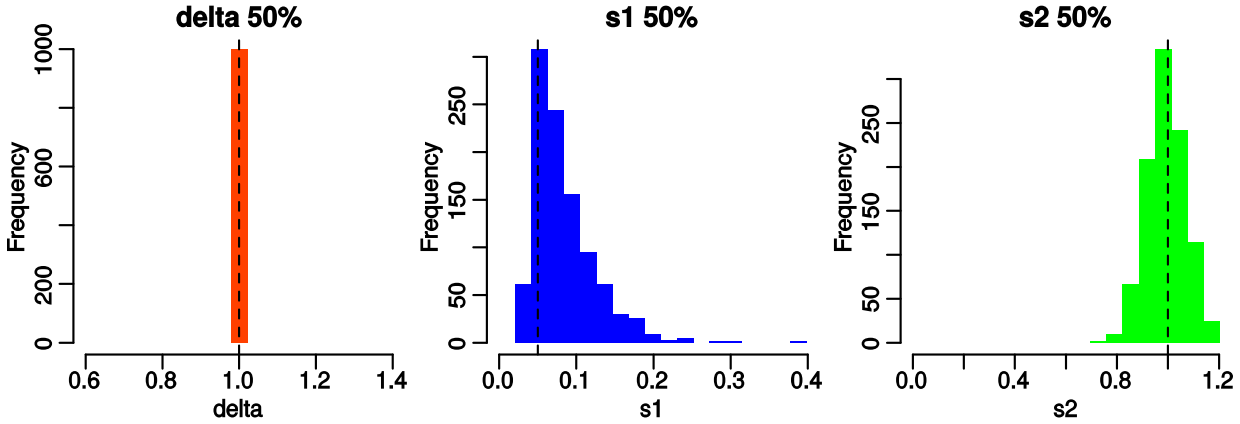
な架空データにあてはめる



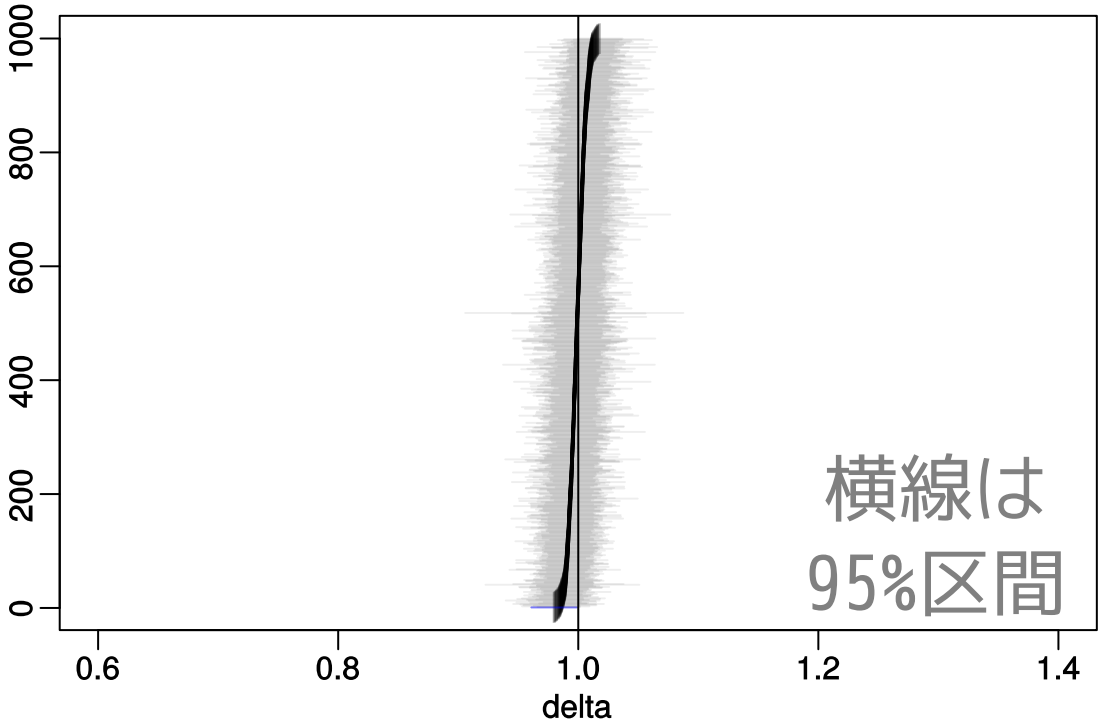


「傾き」δの事後分布を見る



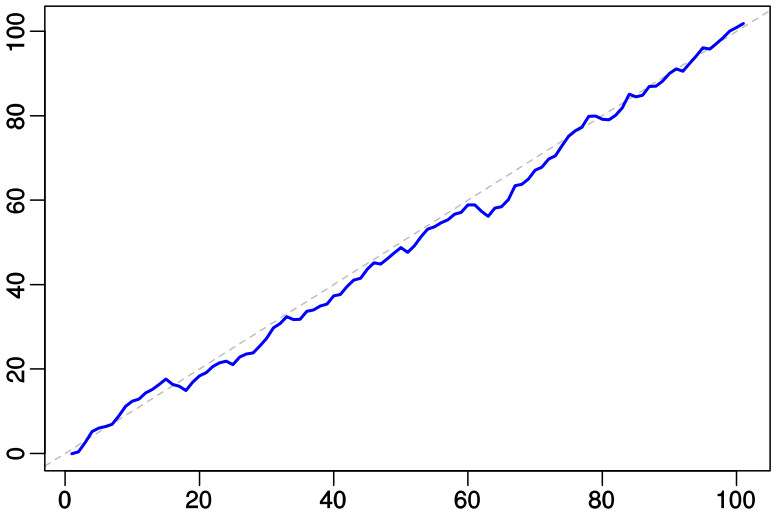


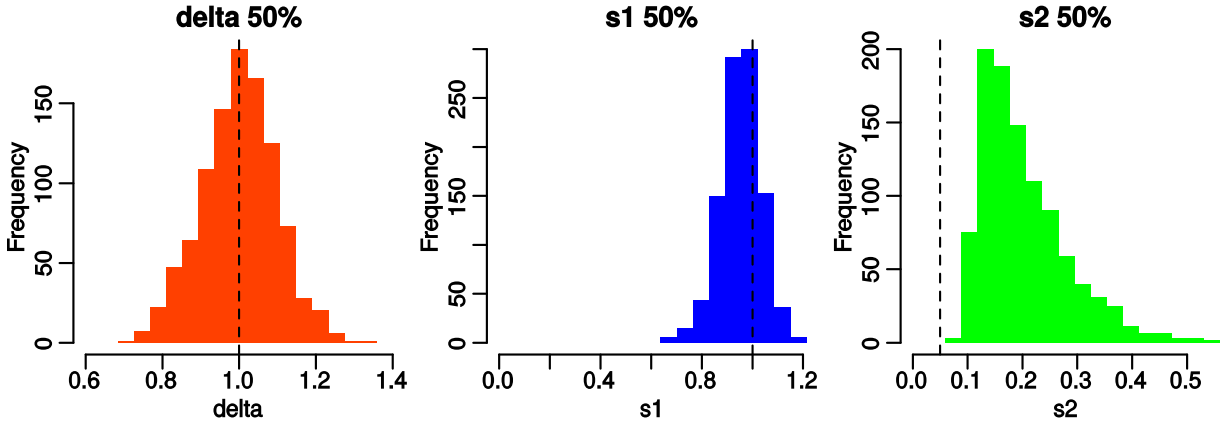
真のδは 1



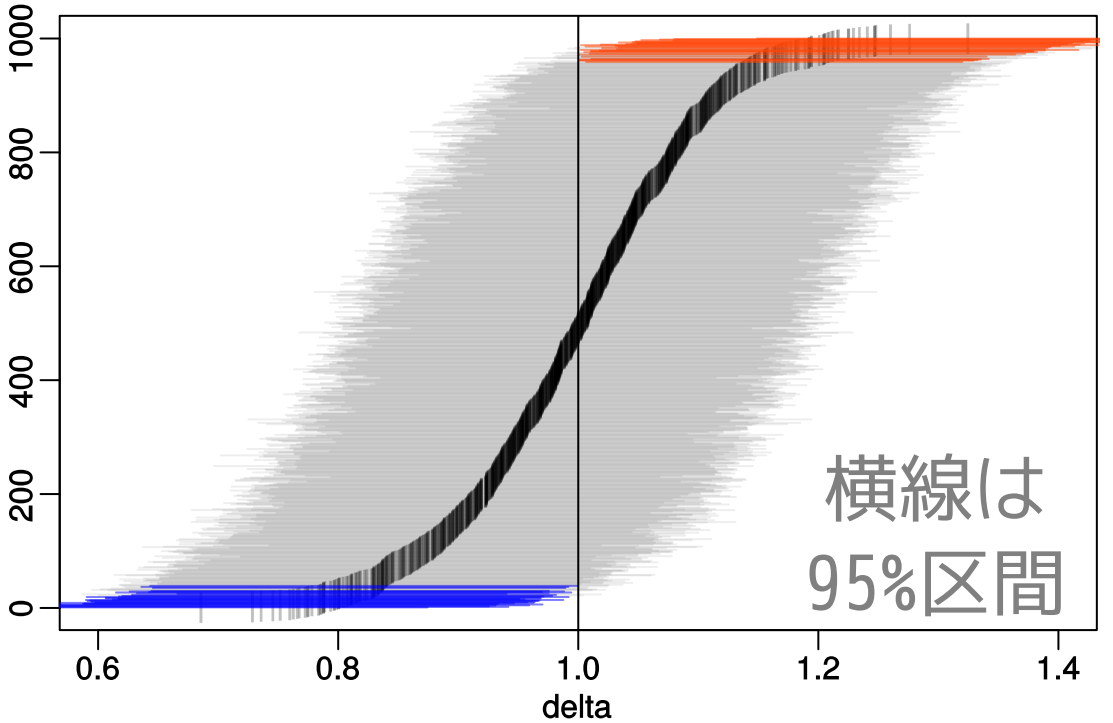
1000回中
1回ずれた

「傾き」δの事後分布を見る



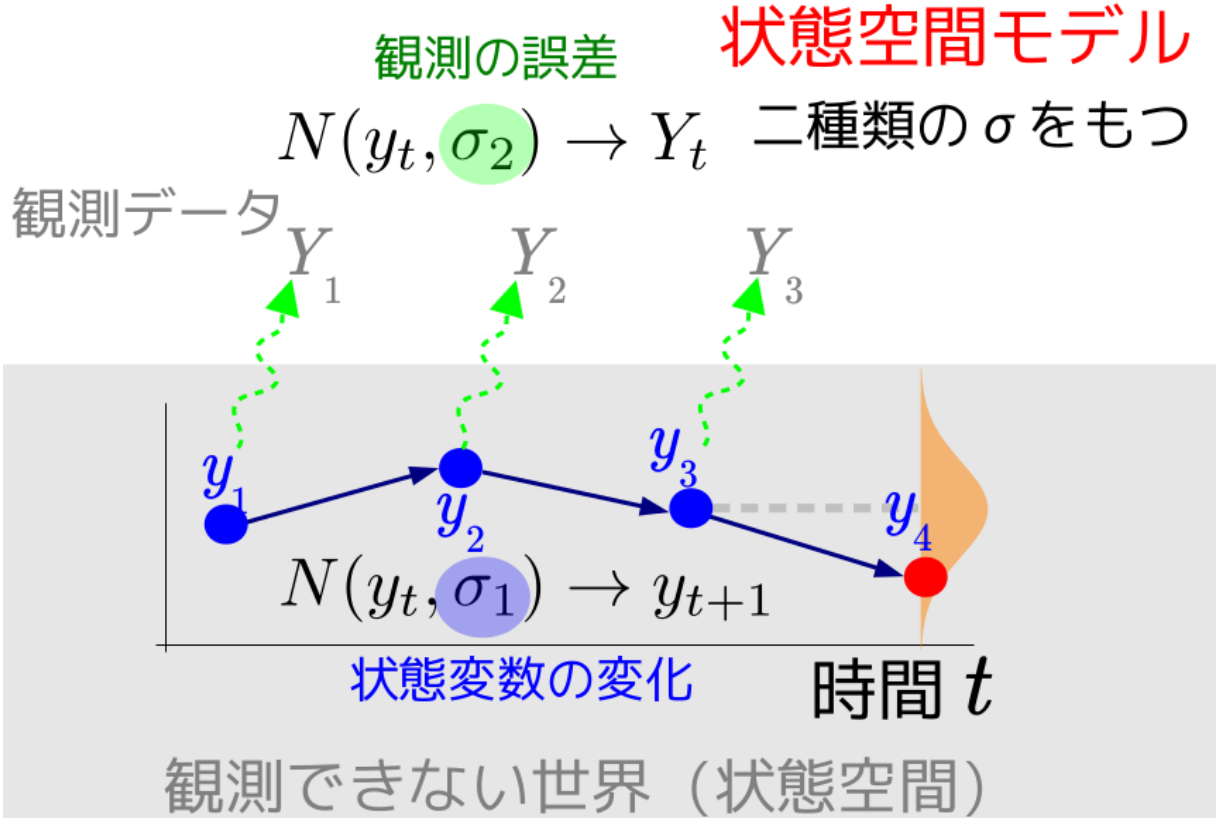


真のδは 1

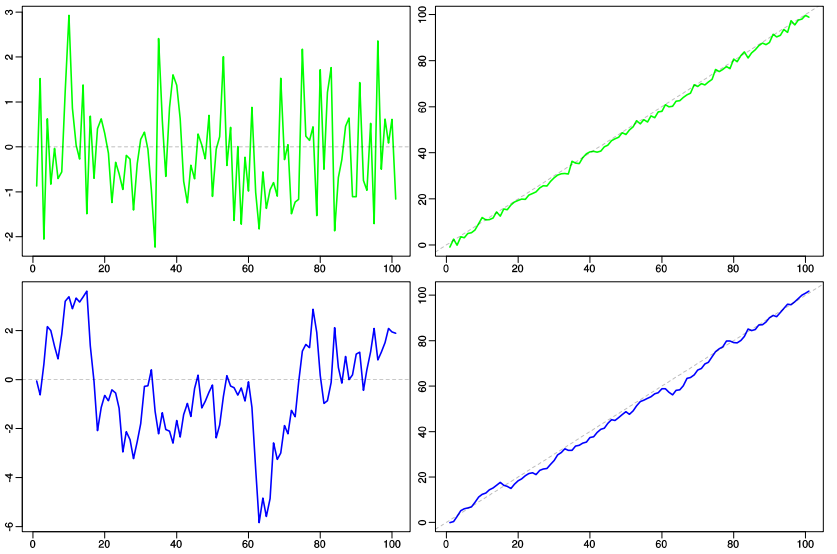


1000回中
62回ずれた

とりあえずの結論



ひとつの状態空間モデルを使って



右の4状態は区別可能でしょう

（危 2）時系列データ X_t

と 時系列データ Y_t

$Y_t \sim X_t$ なうたがわしい回帰

spurious regression

時間があればデモ

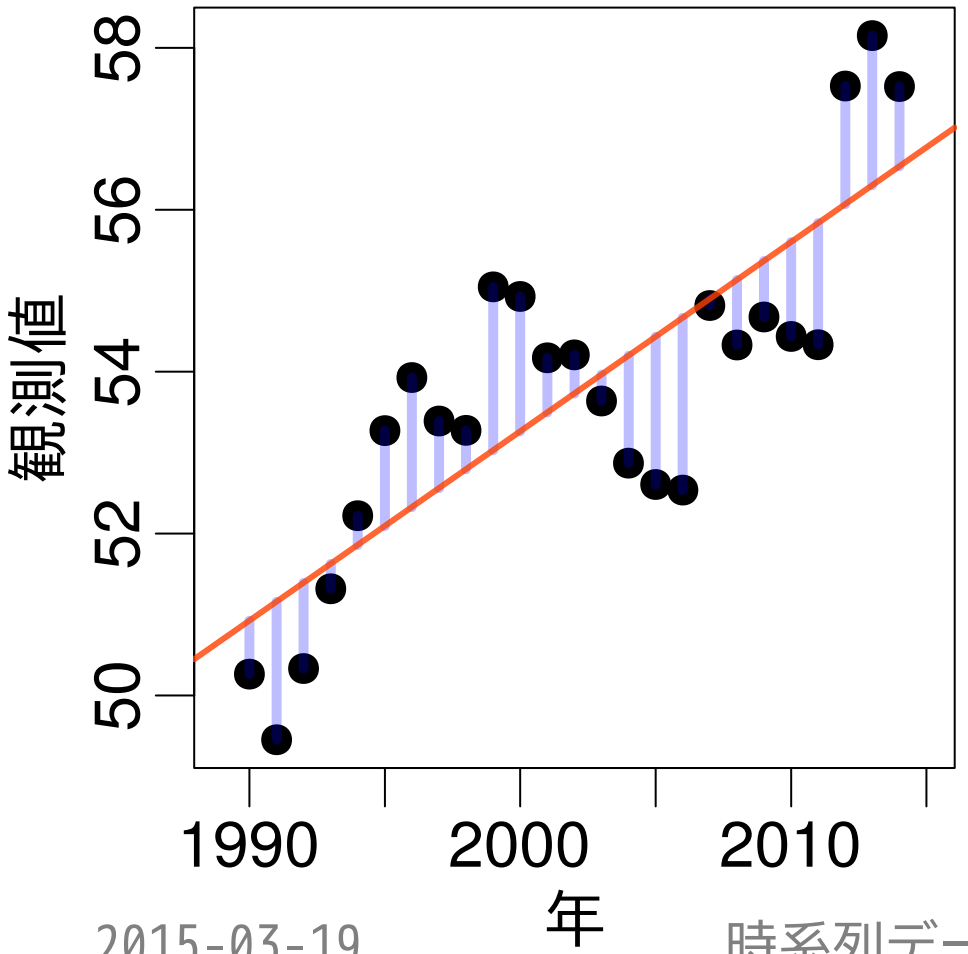
（すみません）

時間的な相関はデータの

情報量を減少させる

空間相関も…

時系列の「ずれ」



GLM のずれ