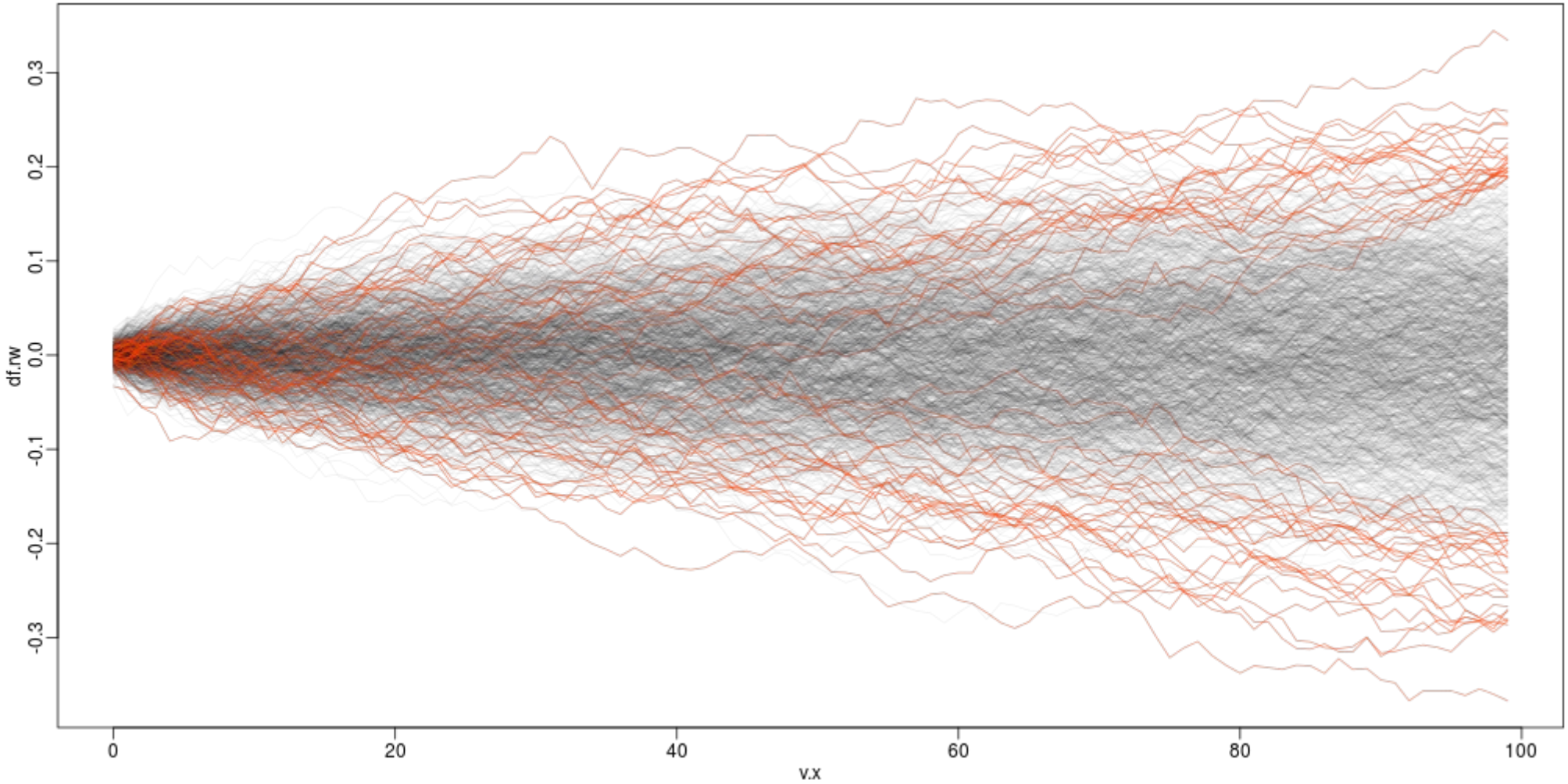
時系列データ解析でよく見る『あぶない』モデリング

久保拓弥（北海道大・環境科学）



今日の要点

「あぶない」時系列データ解析は

やめましょう!

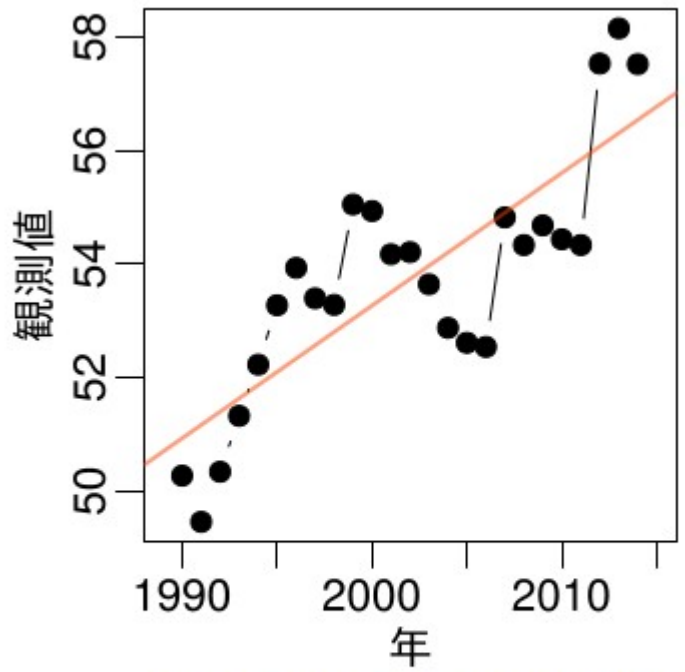
統計モデル
のあてはめ

(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?

（危 2）時系列 Y_t ～ 時系列 X_t

「相関は因果関係ではない」

問題の一部：**にせの回帰**

spurious_regression.R

Source on Save | Run | Source

```
1 x <- cumsum(rnorm(100))
2 y <- cumsum(rnorm(100))
3 plot(ts(x), col = "blue", ylim = range(x, y))
4 lines(ts(y), col = "red")
5 print(summary(glm(y ~ x))$coefficients)
```

5:40 (Top Level) R Script

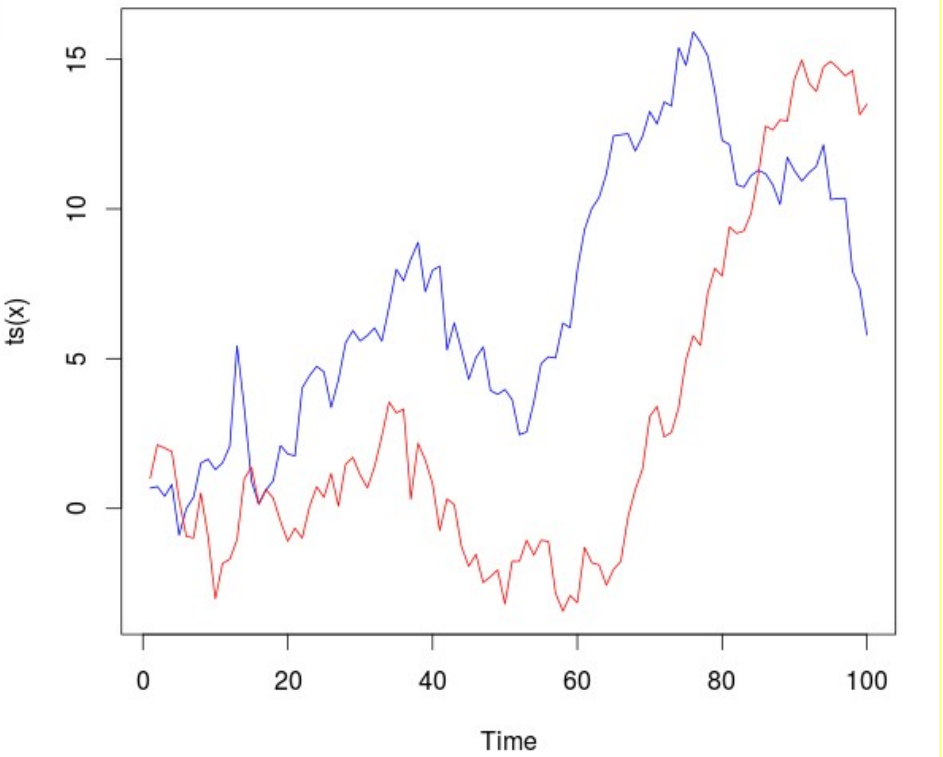
Console

```
> plot(ts(x), col = "blue", ylim = range(x, y))

> lines(ts(y), col = "red")

> print(summary(glm(y ~ x))$coefficients)
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) -1.67120    0.90288 -1.8510 6.7186e-02
x            0.64551    0.10803  5.9753 3.7127e-08
>
```

Environment | History

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish

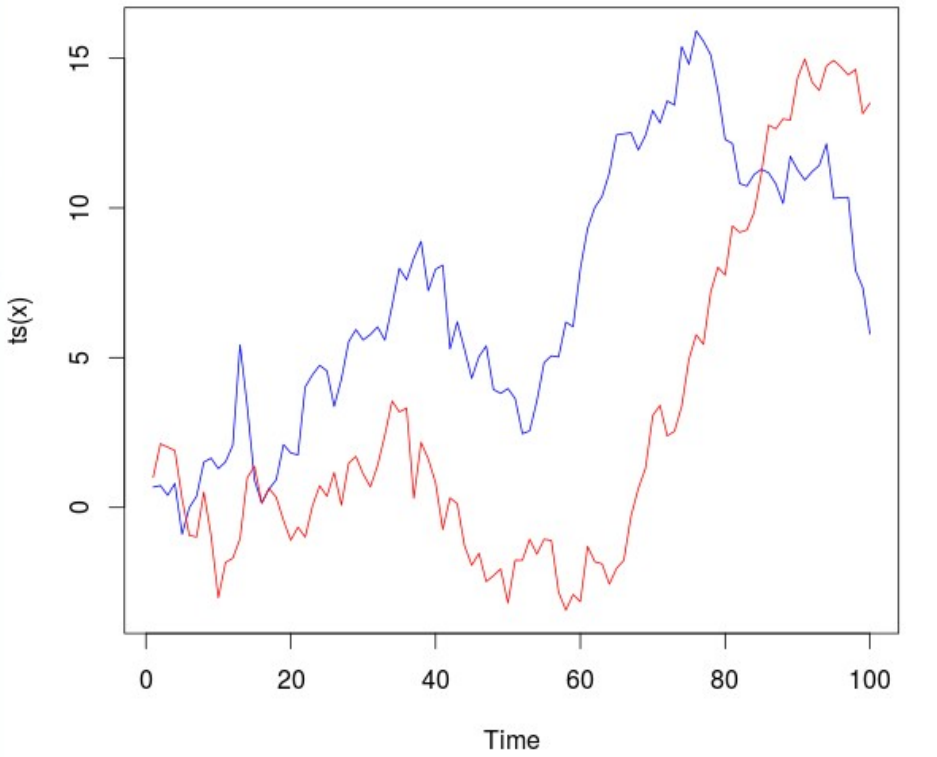


「見せかけの回帰」spurious regression

spurious_regression.R

```
1 x <- cumsum(rnorm(100))
2 y <- cumsum(rnorm(100))
3 plot(ts(x), col = "blue", ylim = range(x, y))
4 lines(ts(y), col = "red")
5 print(summary(glm(y ~ x))$coefficients)
```

```
> plot(ts(x), col = "blue", ylim = range(x, y))

> lines(ts(y), col = "red")

> print(summary(glm(y ~ x))$coefficients)
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) -1.67120    0.90288 -1.8510 6.7186e-02
x            0.64551    0.10803  5.9753 3.7127e-08
```



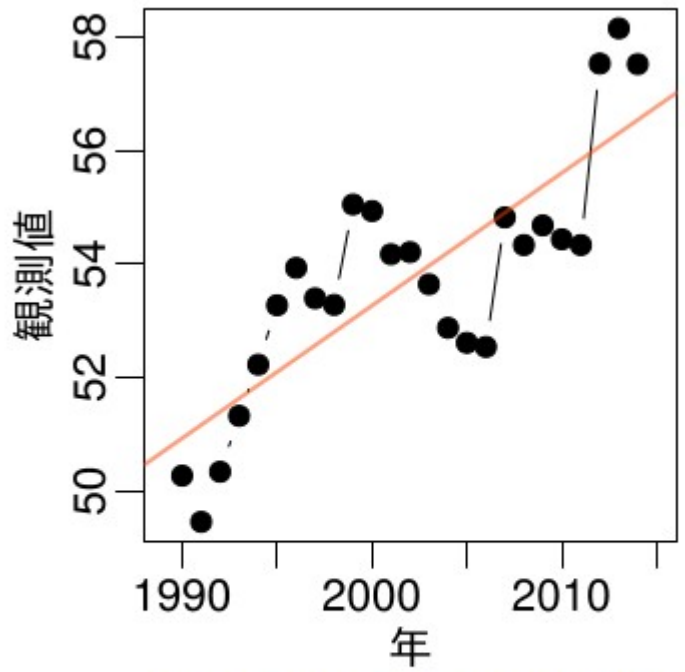
ちょっとだけ実演してみます

時系列データの統計モデリング

- 安易に「回帰」してはいけない
- ランダムウォークモデルが基本
- 統計モデルが生成する時系列パターンを意識する
- 階層ベイズモデルで推定

状態空間モデル

（危 1）時系列データを GLM で



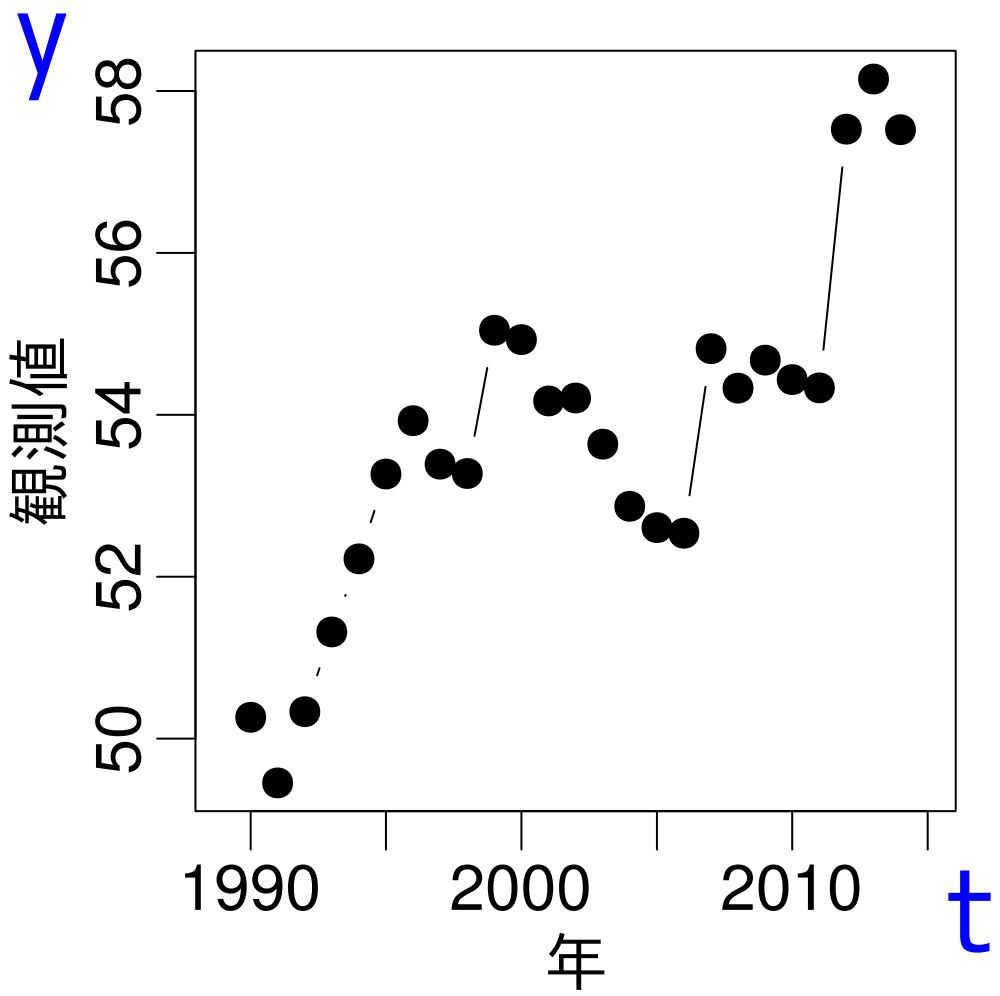
「ゆーいな傾き」を
ねつぞうする原因

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?

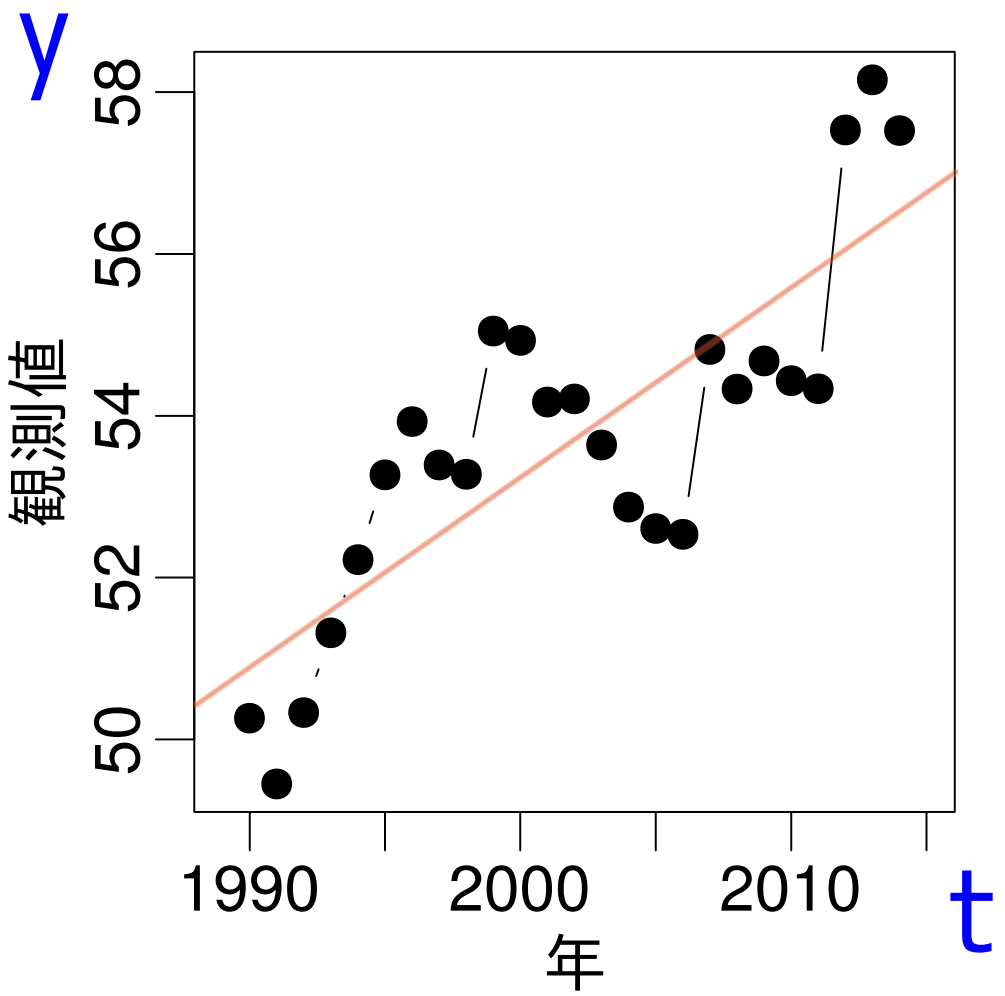
このような時系列データがあったとしましょう



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)

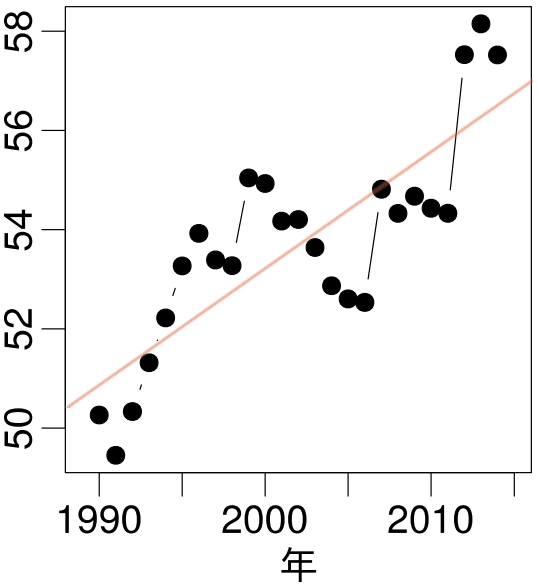
時系列データの統計モデリング入門



glm(y ~ t)

…とモデル
をあてはめてみた

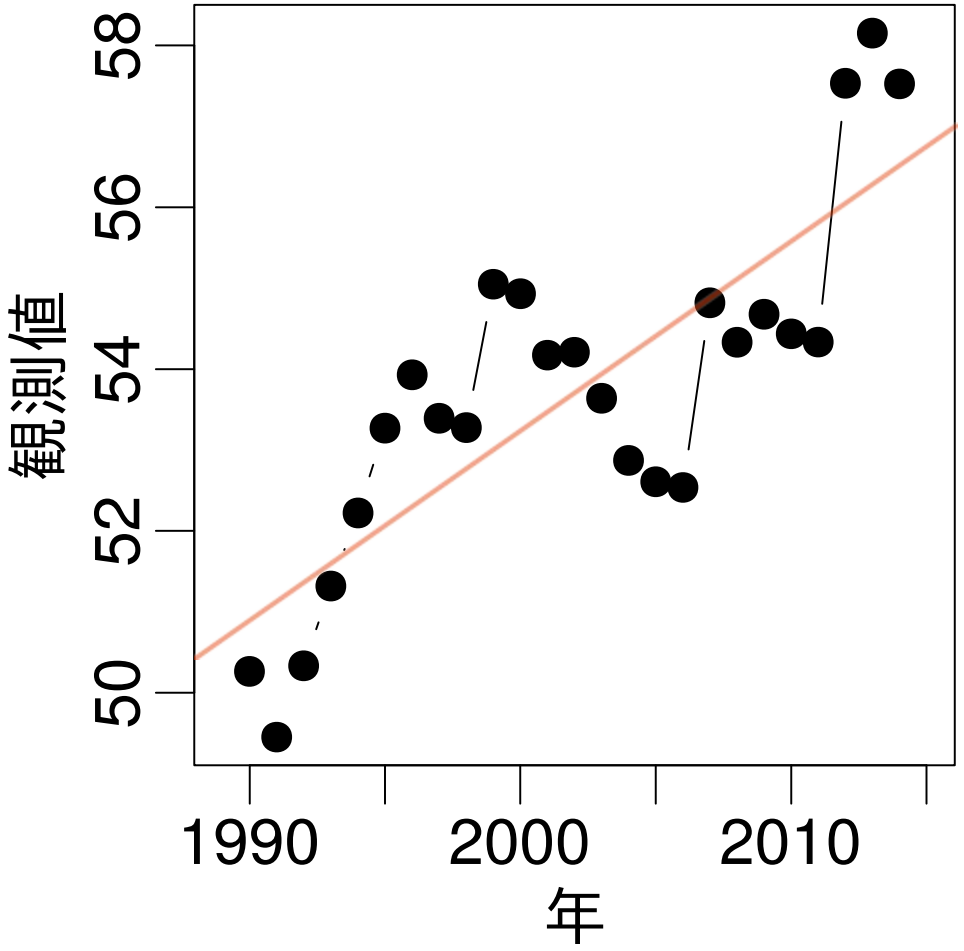
「やったーゆーいだ!!」……??




```
> 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」(偽)

が「ぞろぞろ」でます

傾きの検定やめて

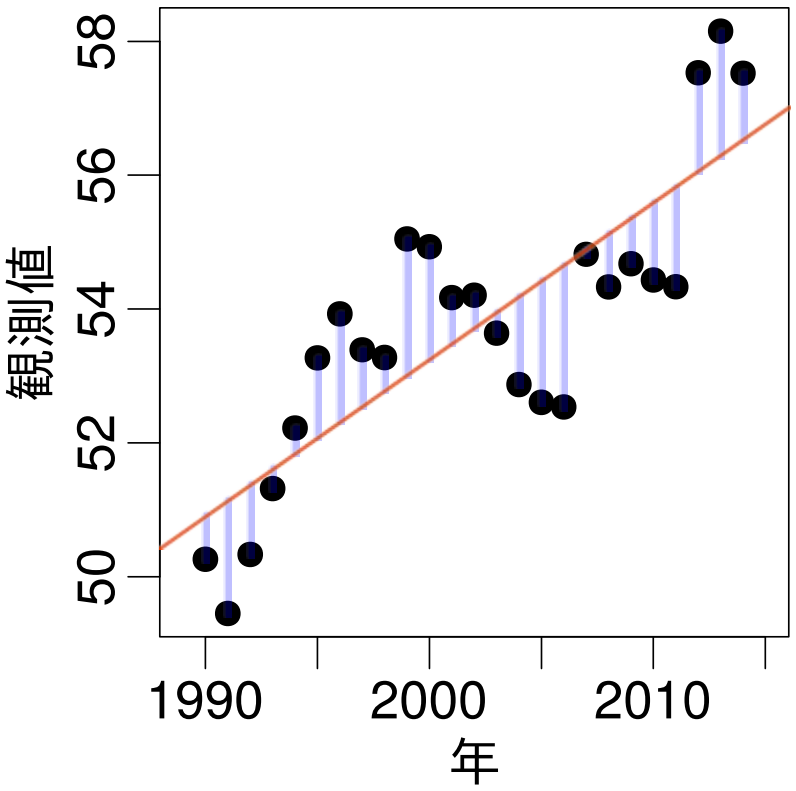
AIC モデル選択

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

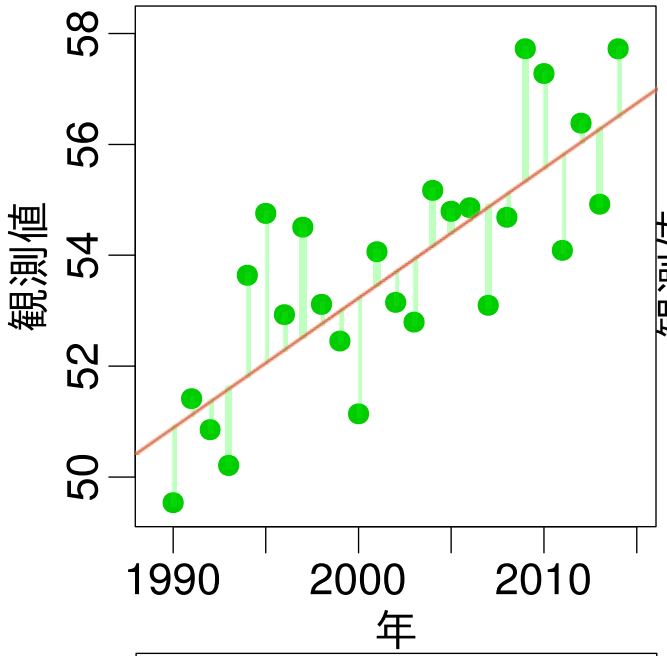
統計モデルがおかしい?

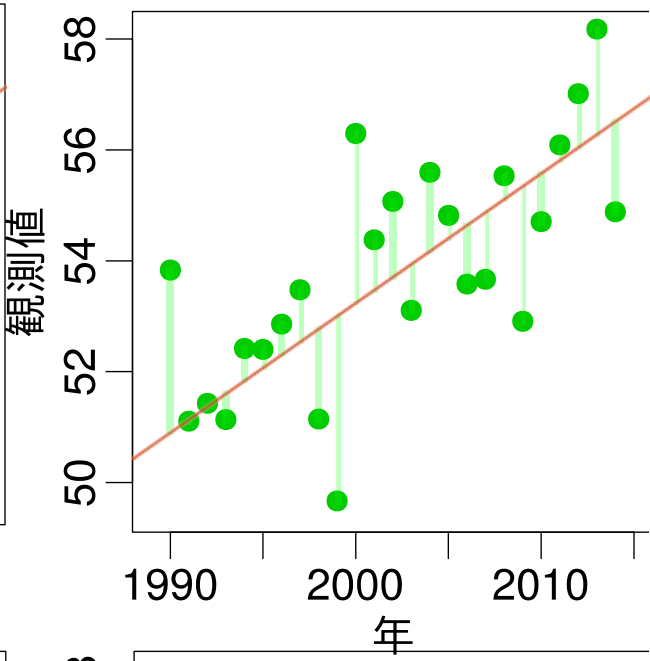
時系列の「ずれ」

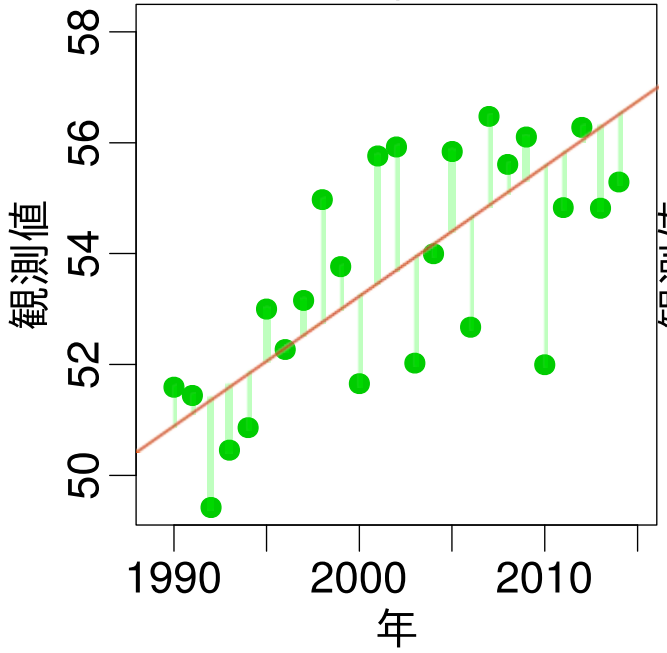


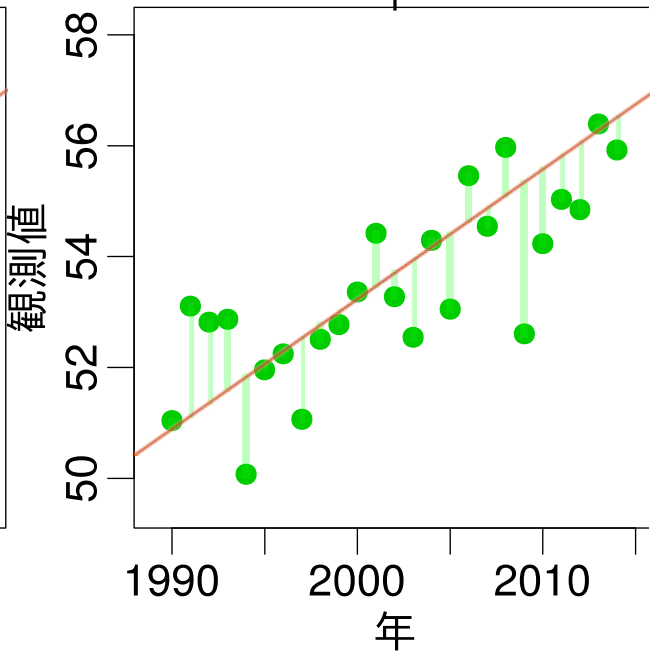
ずれかたが
ちがってる?

GLM のずれ

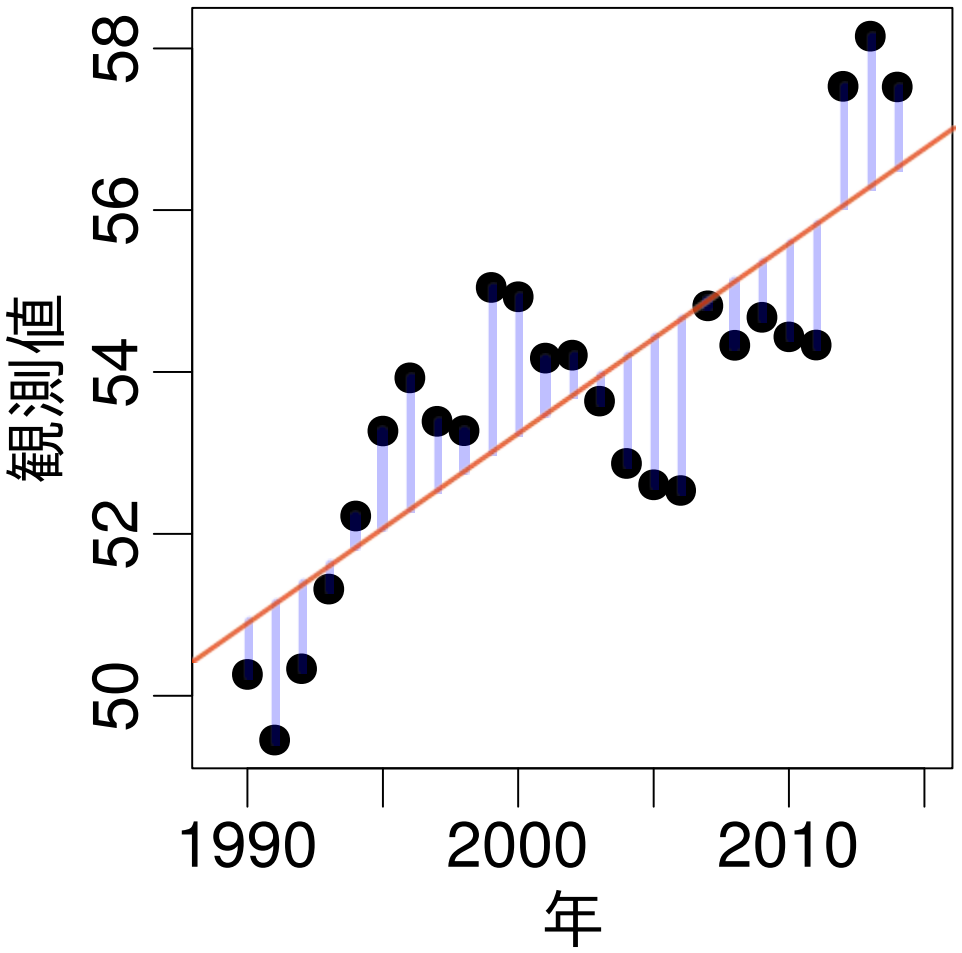




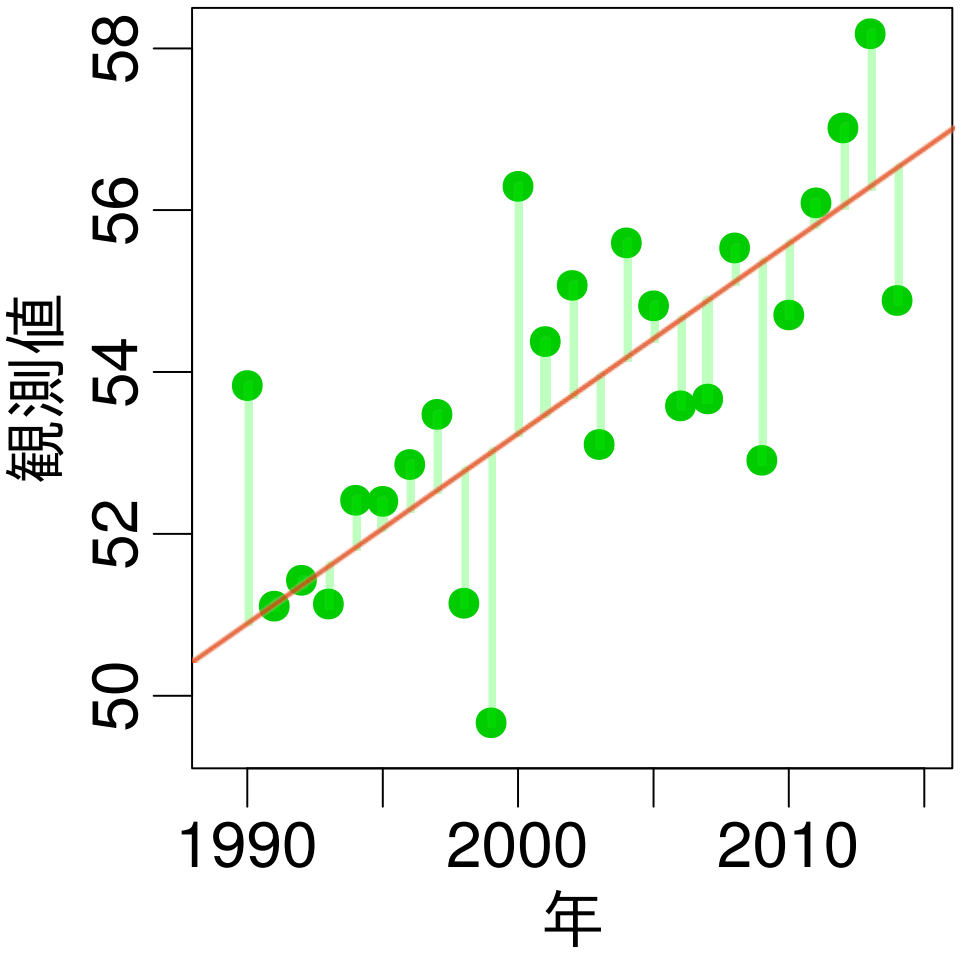




時系列の「ずれ」



GLM のずれ



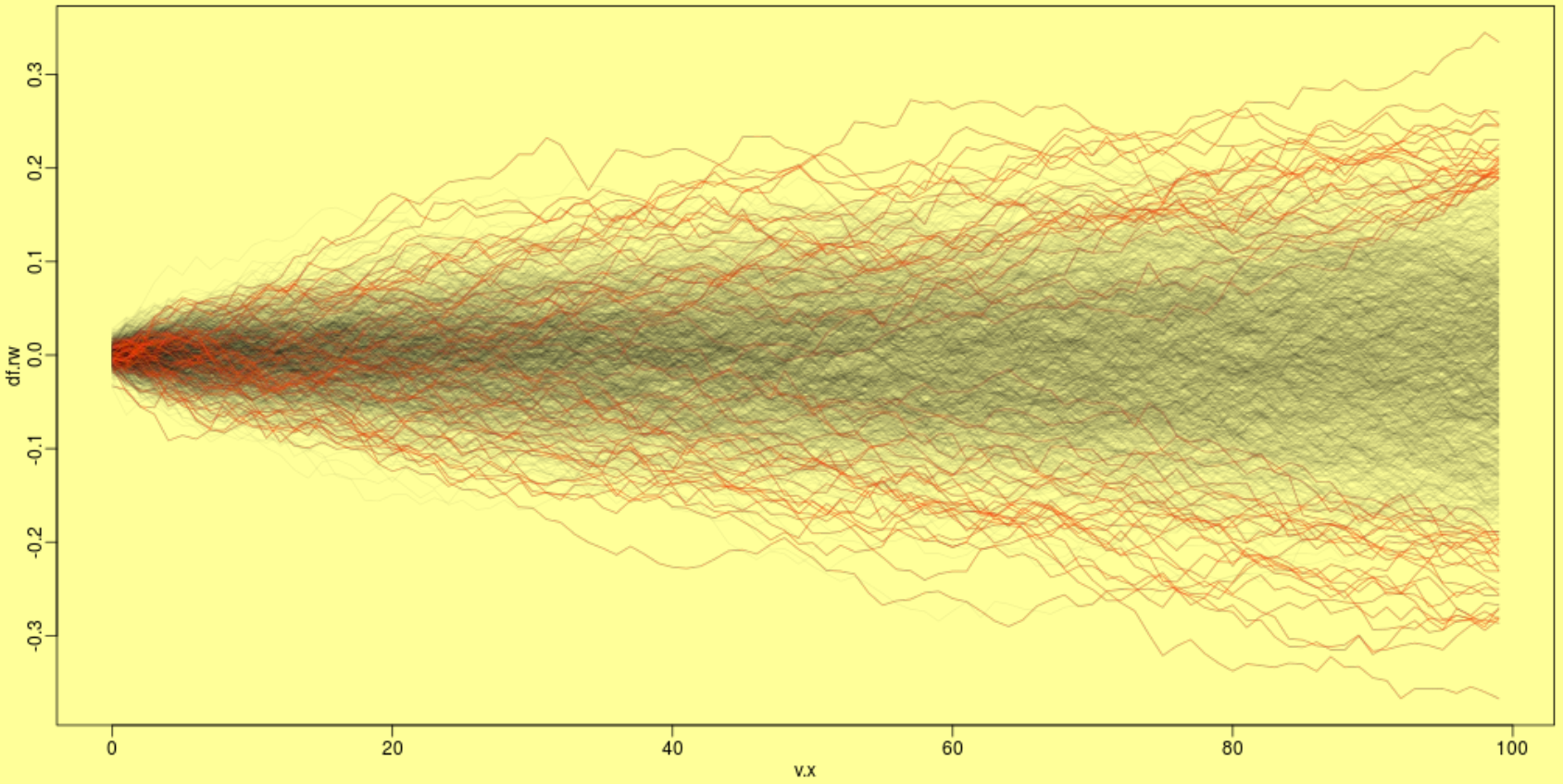
直線からのずれがちがう!

時間的自己相関がある

時間的自己相関がない

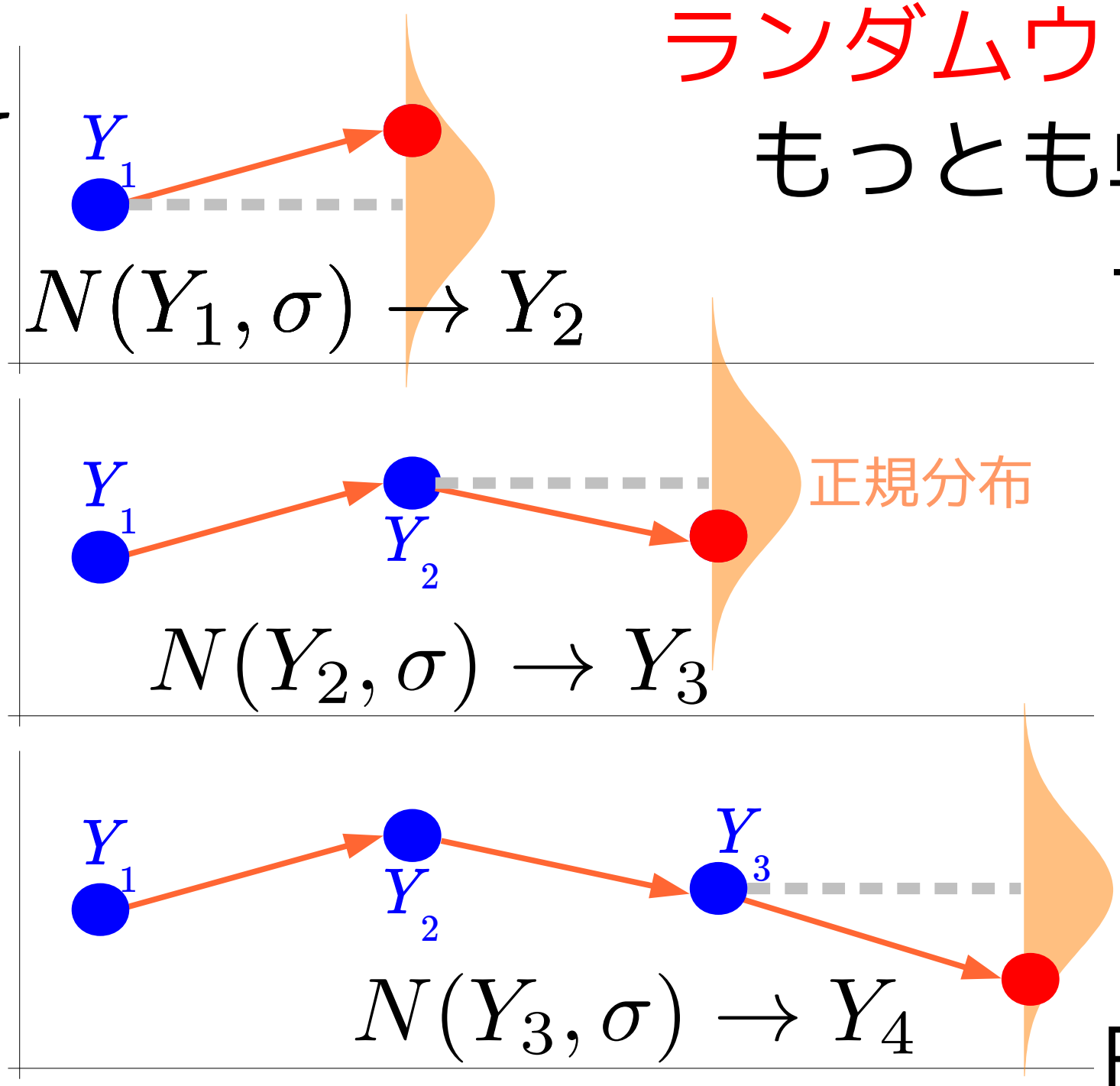
時系列の基本モデルのひとつ

ランダムウォーク（乱歩）



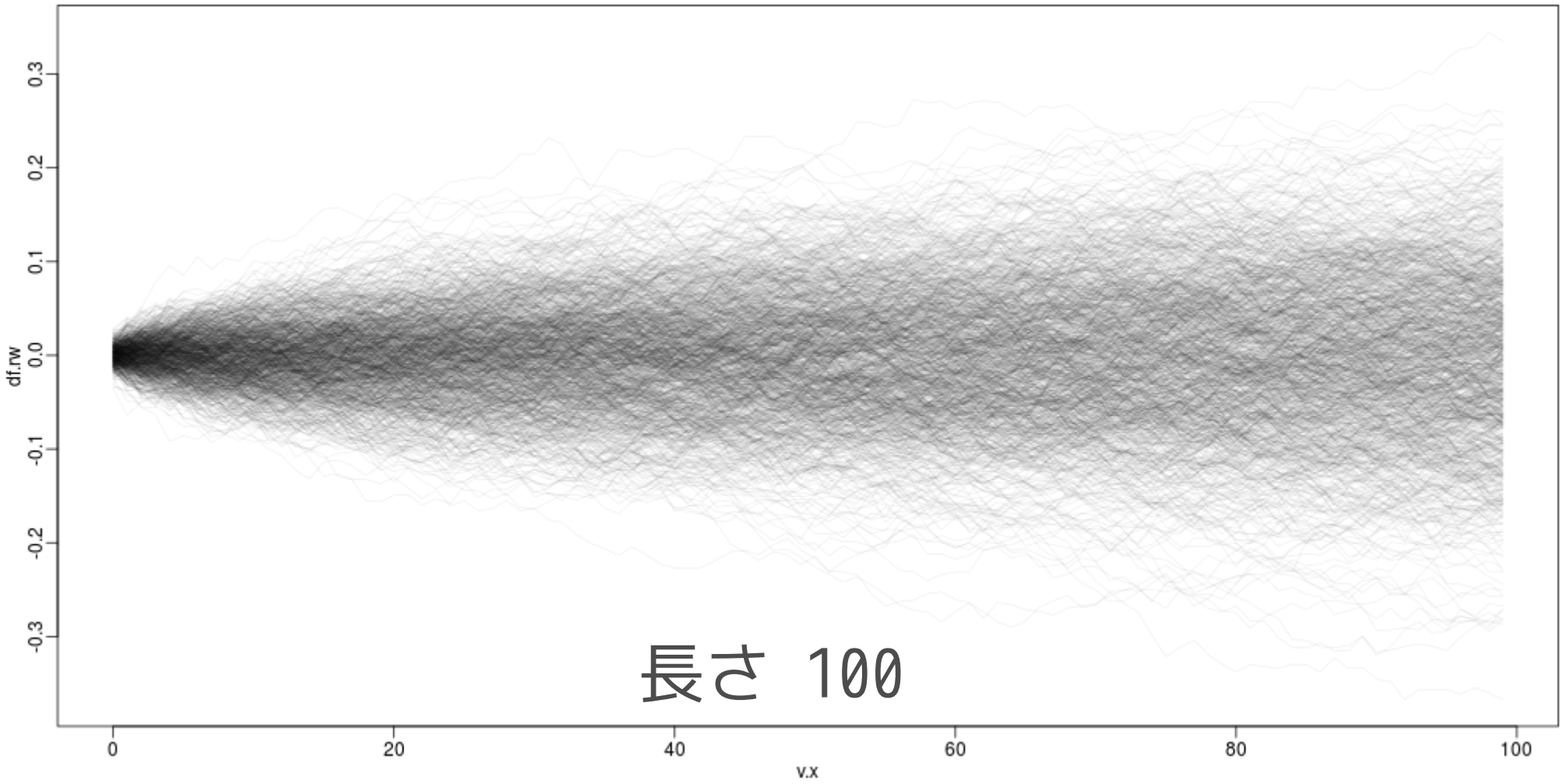
ランダムウォーク

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

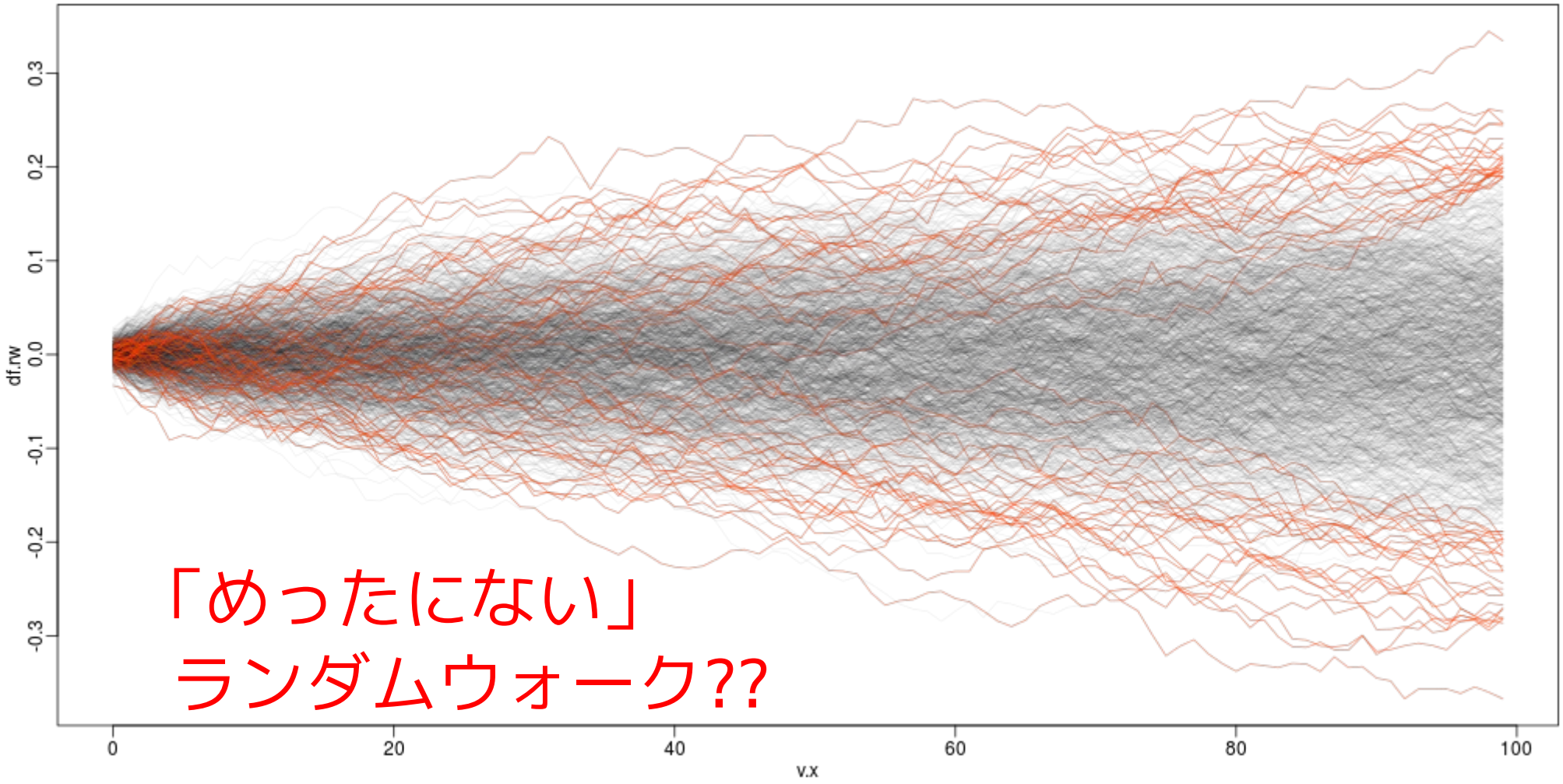
とりあえず 1000 本ほど生成してみました



長さ 100

例外的な時系列というのはありえる

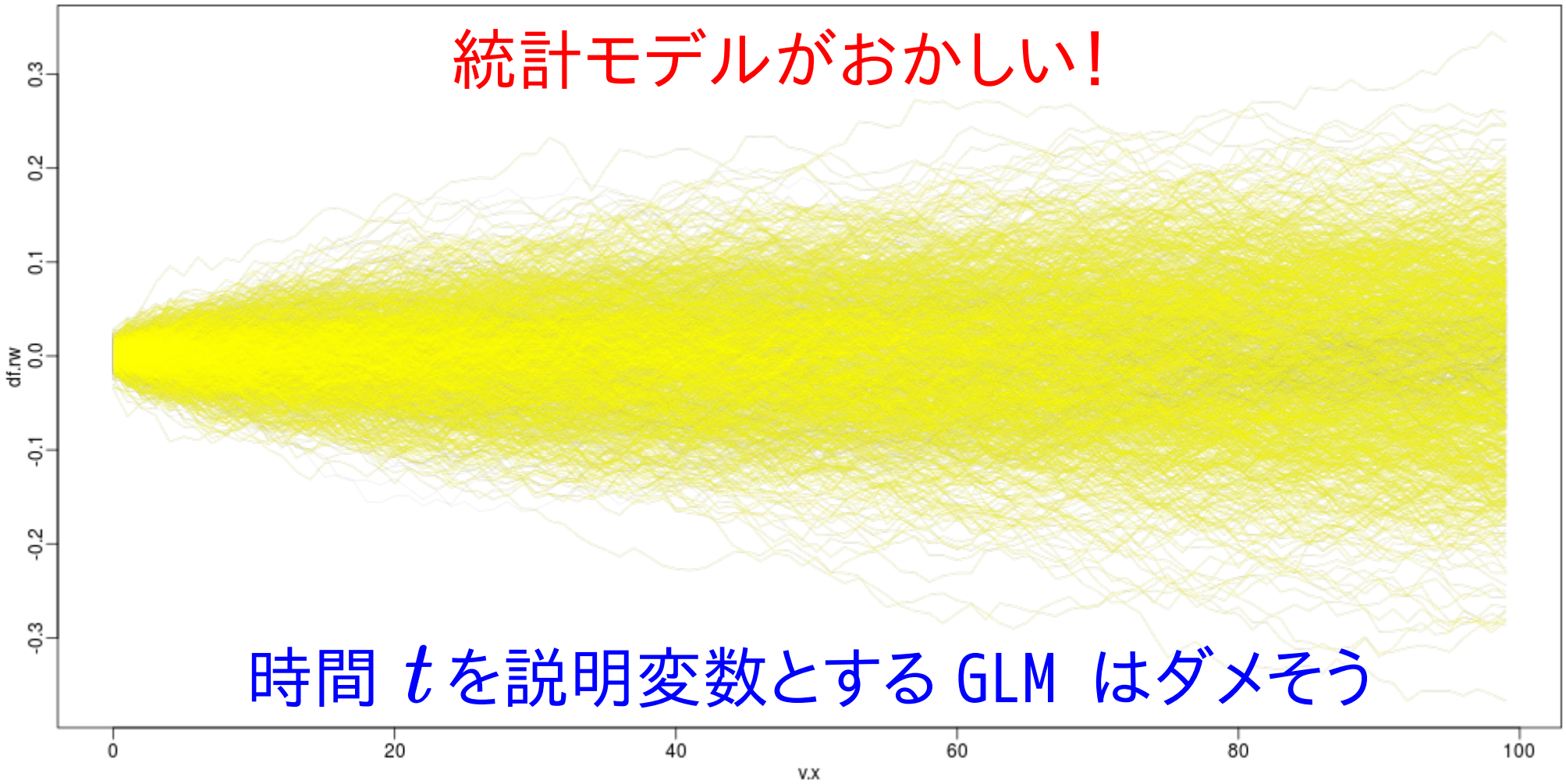
たとえば $t = 100$ でかなり外れている 50 本



「めったにない」
ランダムウォーク??

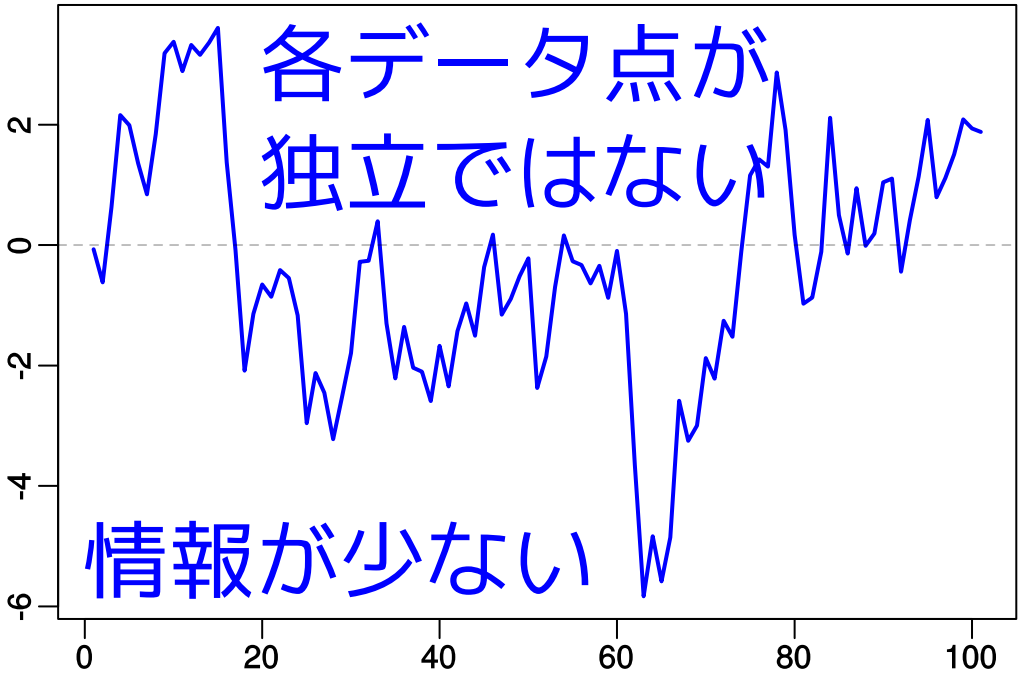
しかし直線回帰 GLM あてはめると…

ほとんどすべての場合で「ゆーい」！

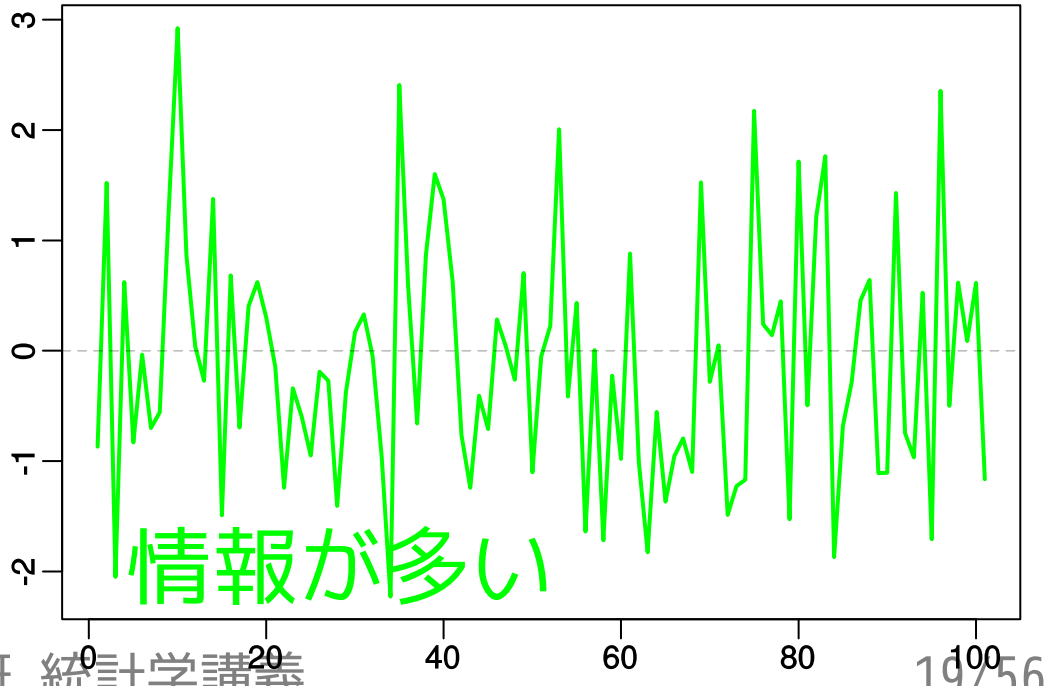


ちょっとでも傾いてたら「ゆーい」

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）

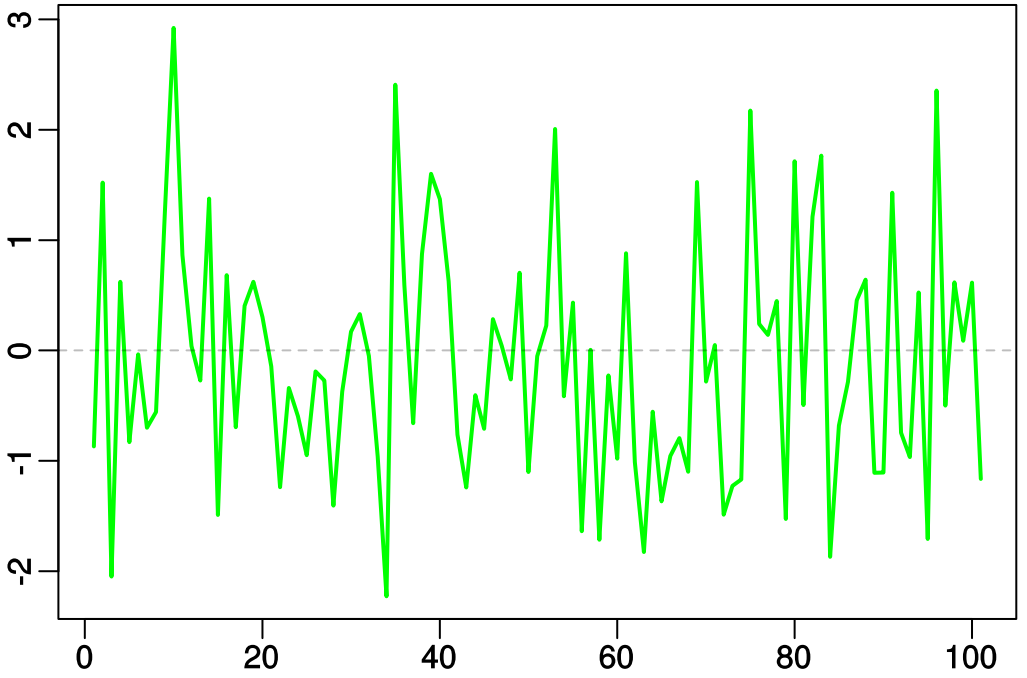
を調べたらいいの?

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

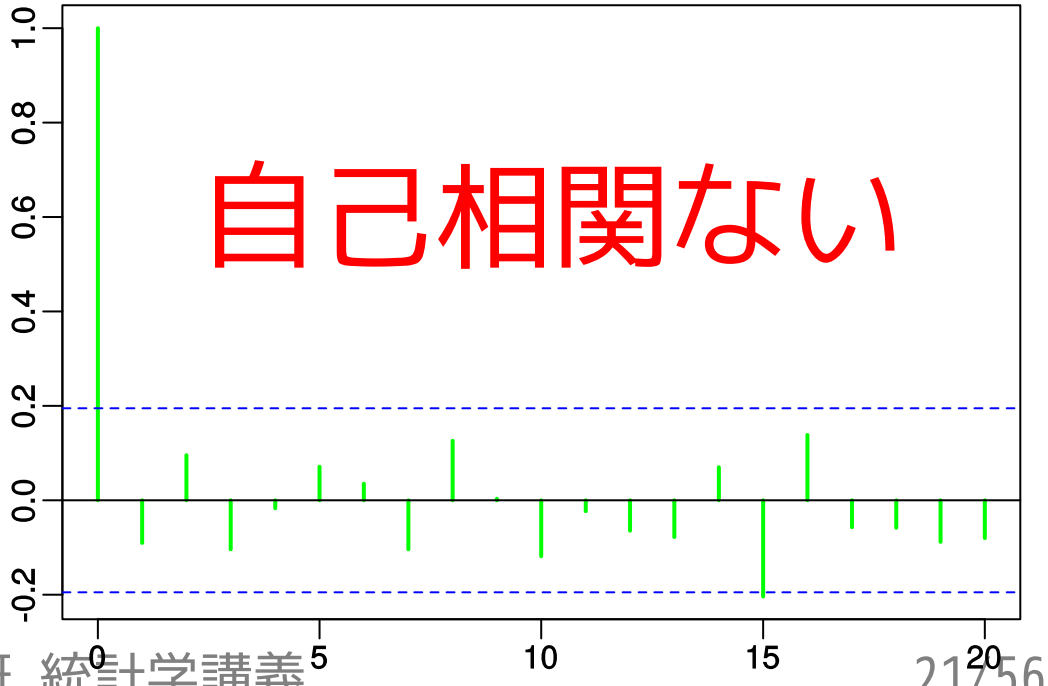
R の ts クラス: 時系列をあつかう

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

これはたんなる
100 個の正規乱数



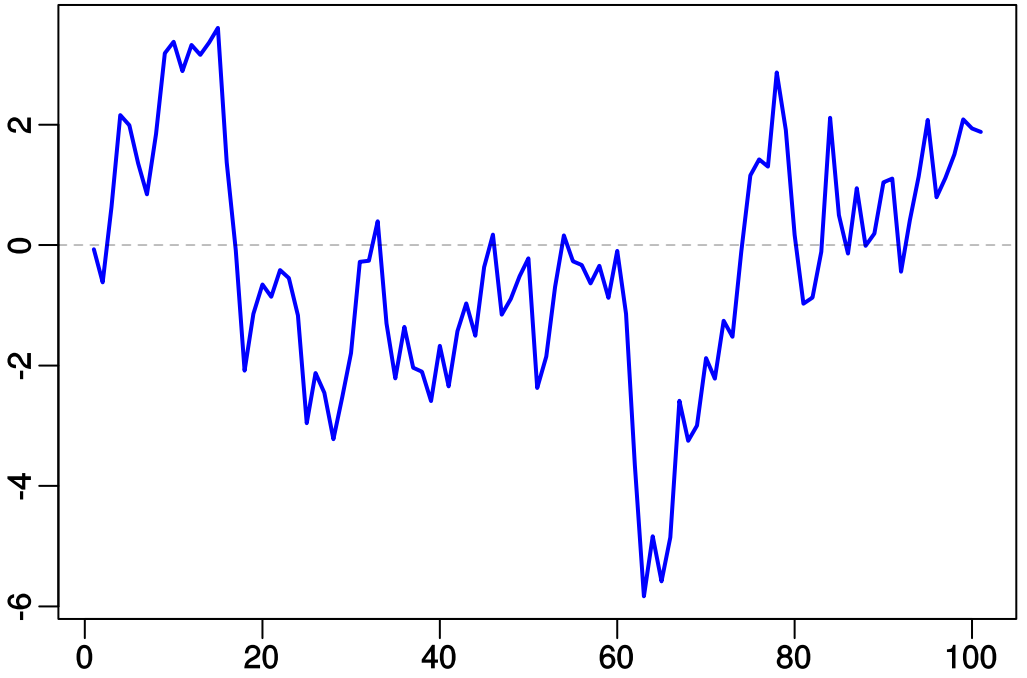

```
plot(acf(ts(Y)))
```



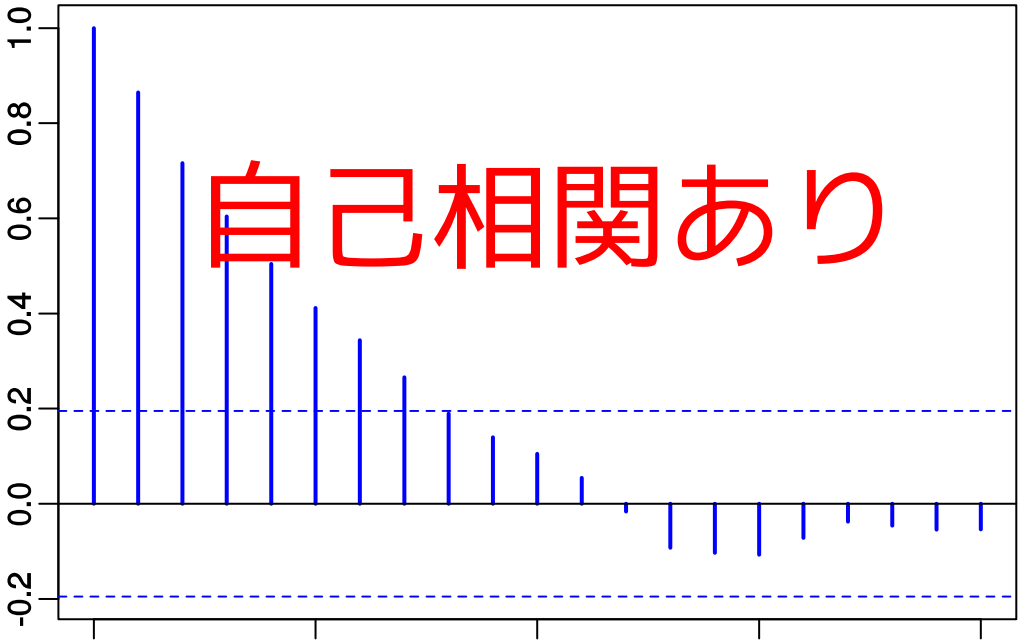
自己相関ない

自己相関減衰の様子を図示

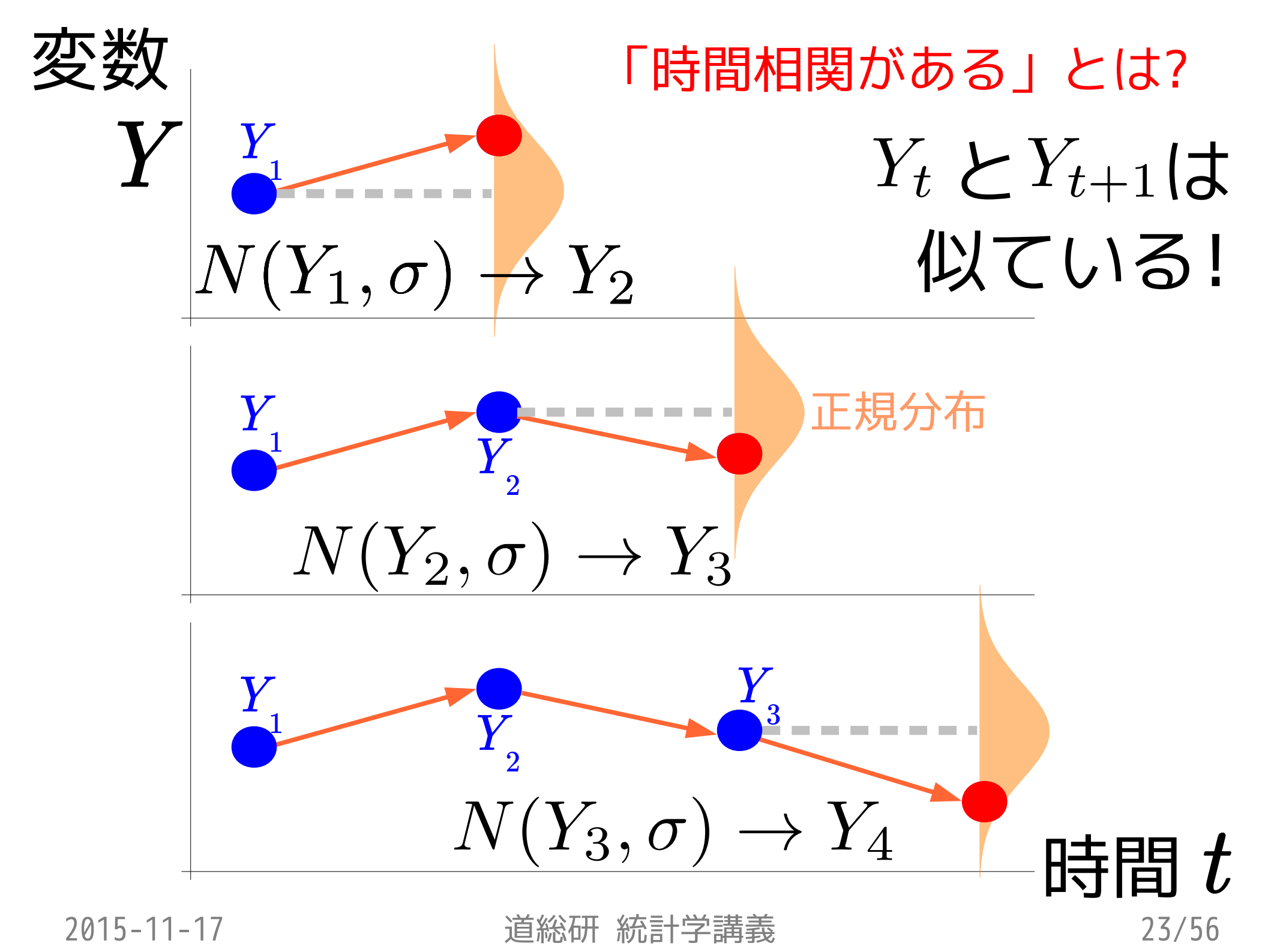
```
plot(ts(Y))
```




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



自己相関あり


変数
Y
「時間相関がある」とは?
Y_t と Y_{t+1} は
似ている!
Y_1
$N(Y_1, \sigma) \to Y_2$
Y_2
$N(Y_2, \sigma) \to Y_3$
正規分布
Y_3
$N(Y_3, \sigma) \to Y_4$
時間 t

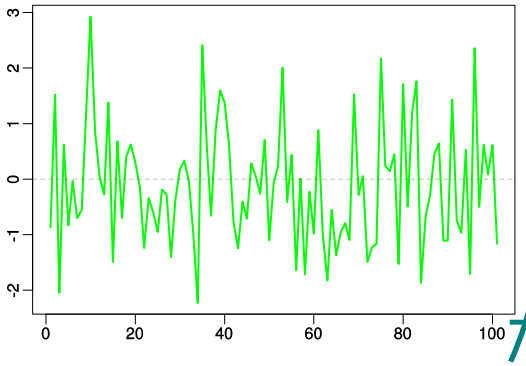
時間的自己相関

はいつも役にたつわけではない?

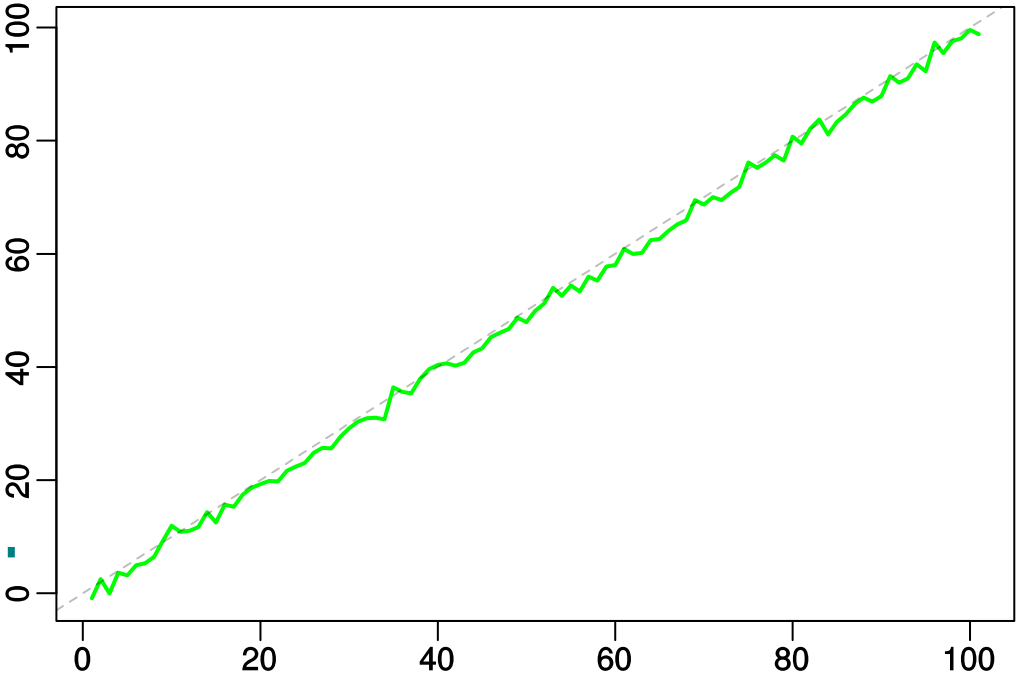
$$\rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-1})}}$$

各点独立のデータをナナメにすると?

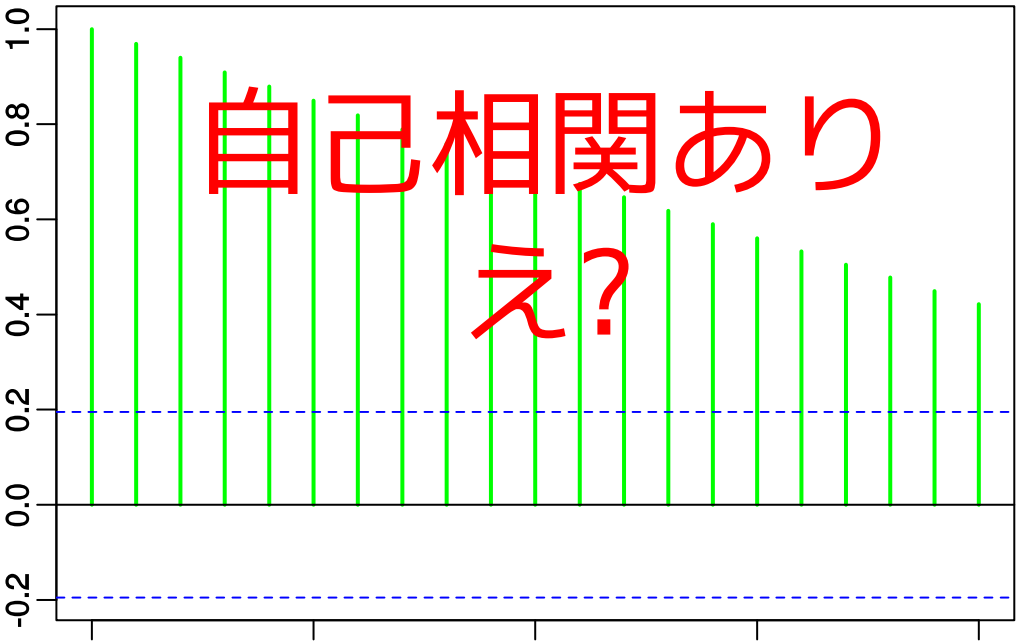
```
plot(ts(Y))
```



これを
ナナメに
したもの
なんだけど…



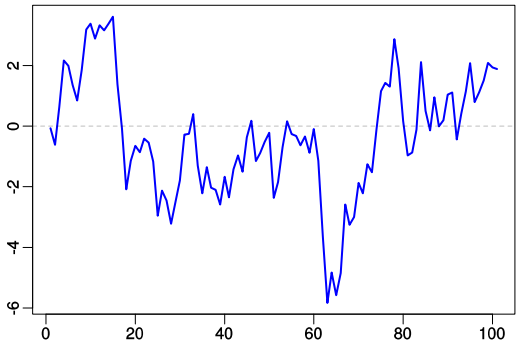

```
plot(acf(ts(Y)))
```



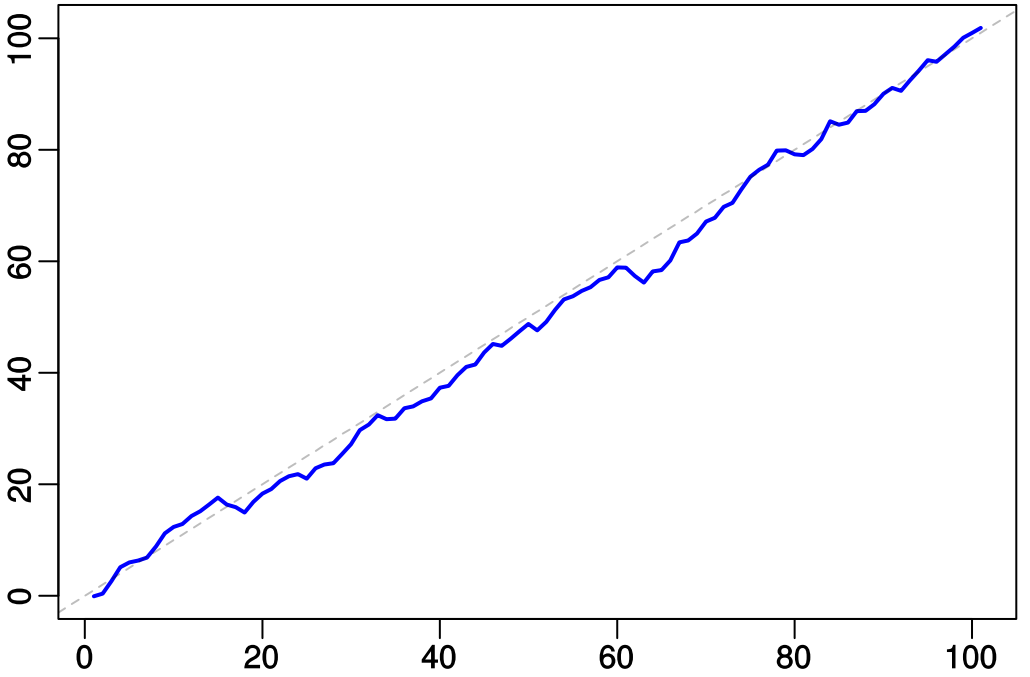
自己相関あり
え?

各点独立のデータをナナメにすると?

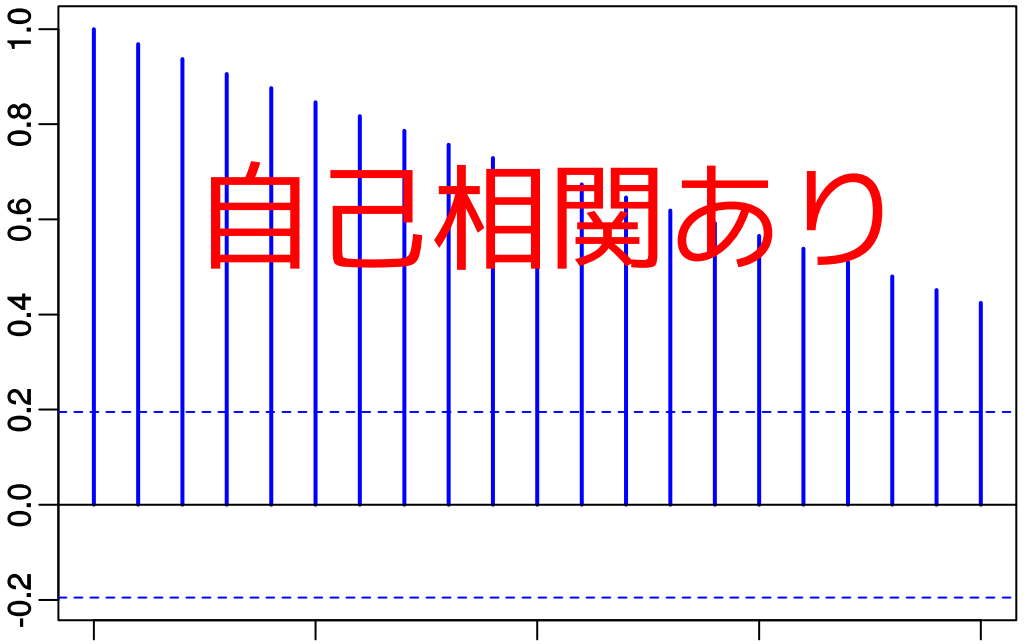
```
plot(ts(Y))
```



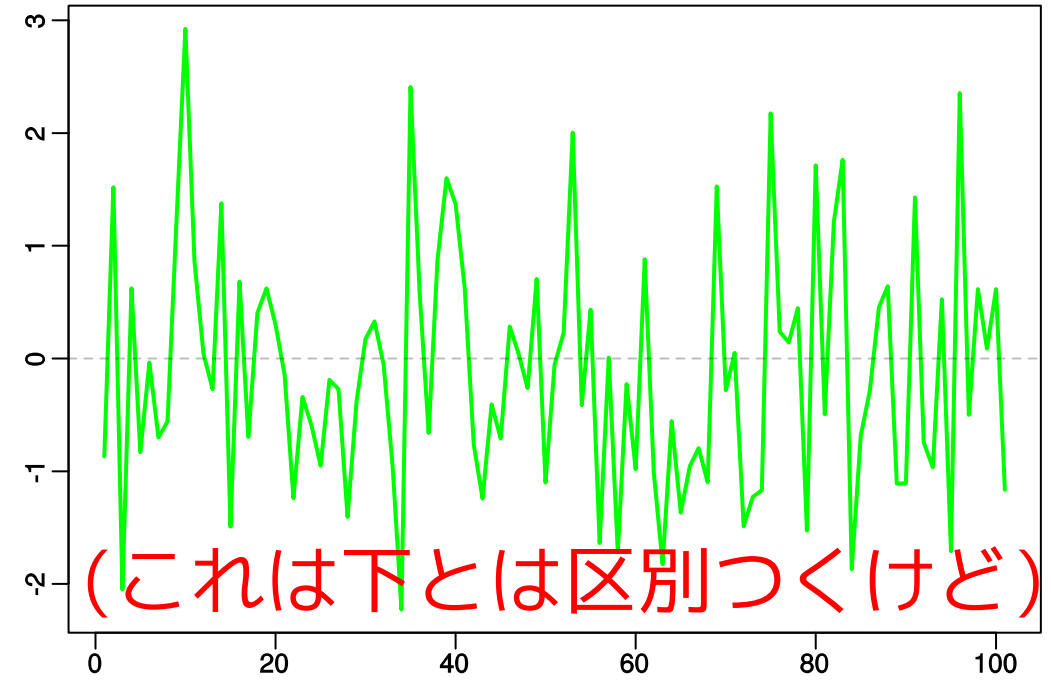
これを
ナナメに
したもの

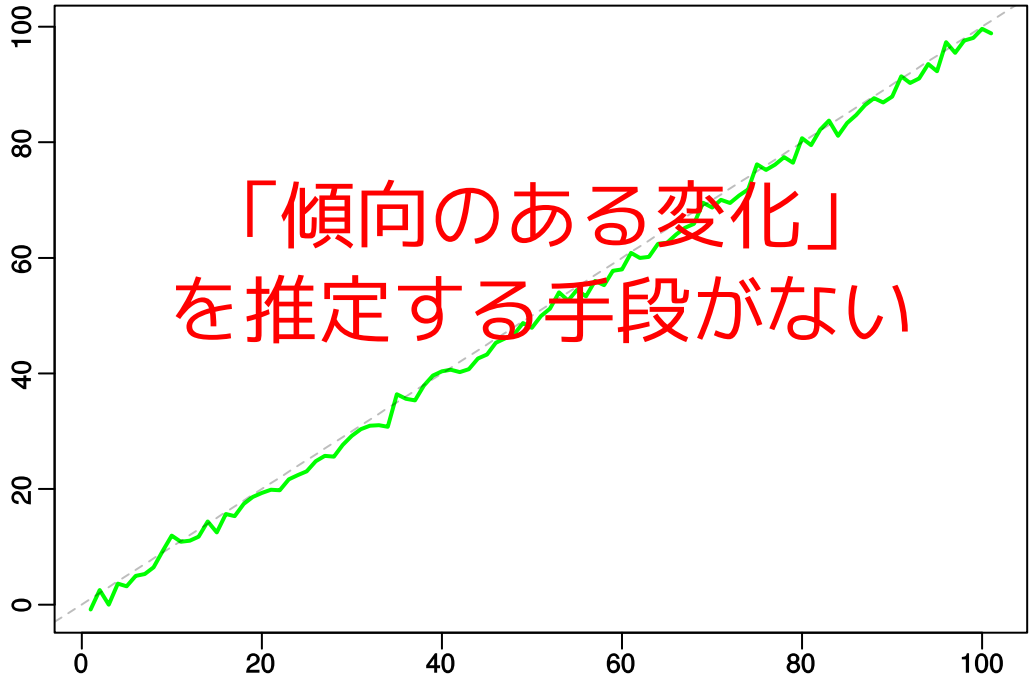


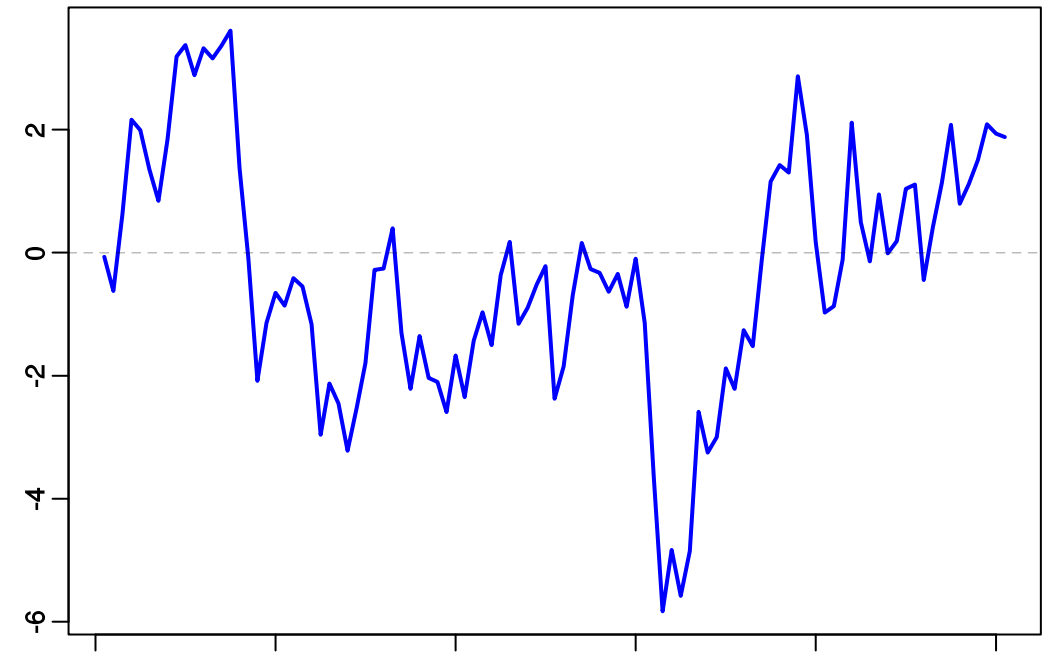

```
plot(acf(ts(Y)))
```

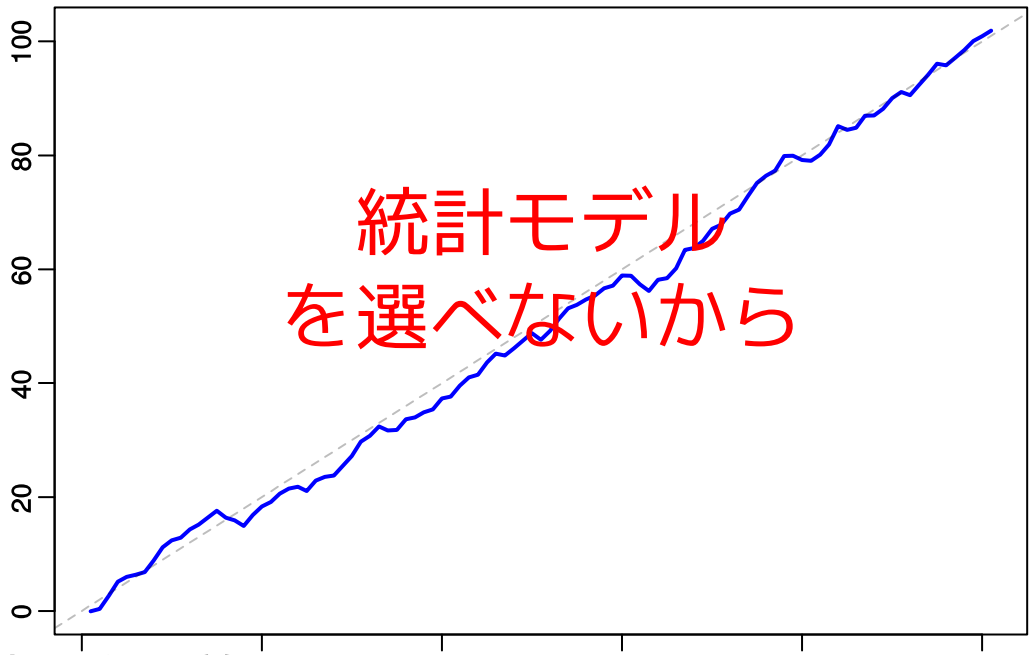


自己相関係数みても区別がつかない



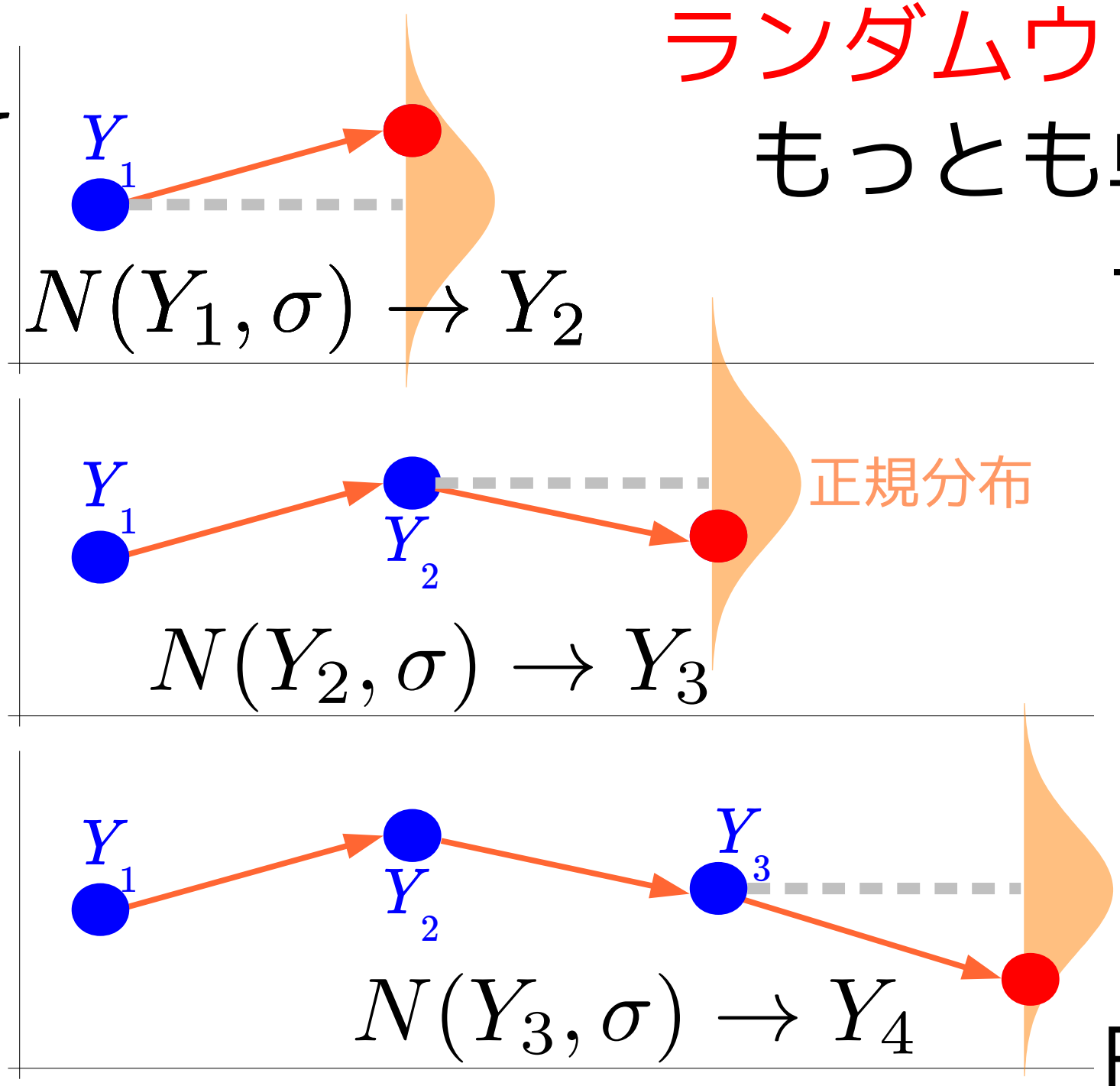






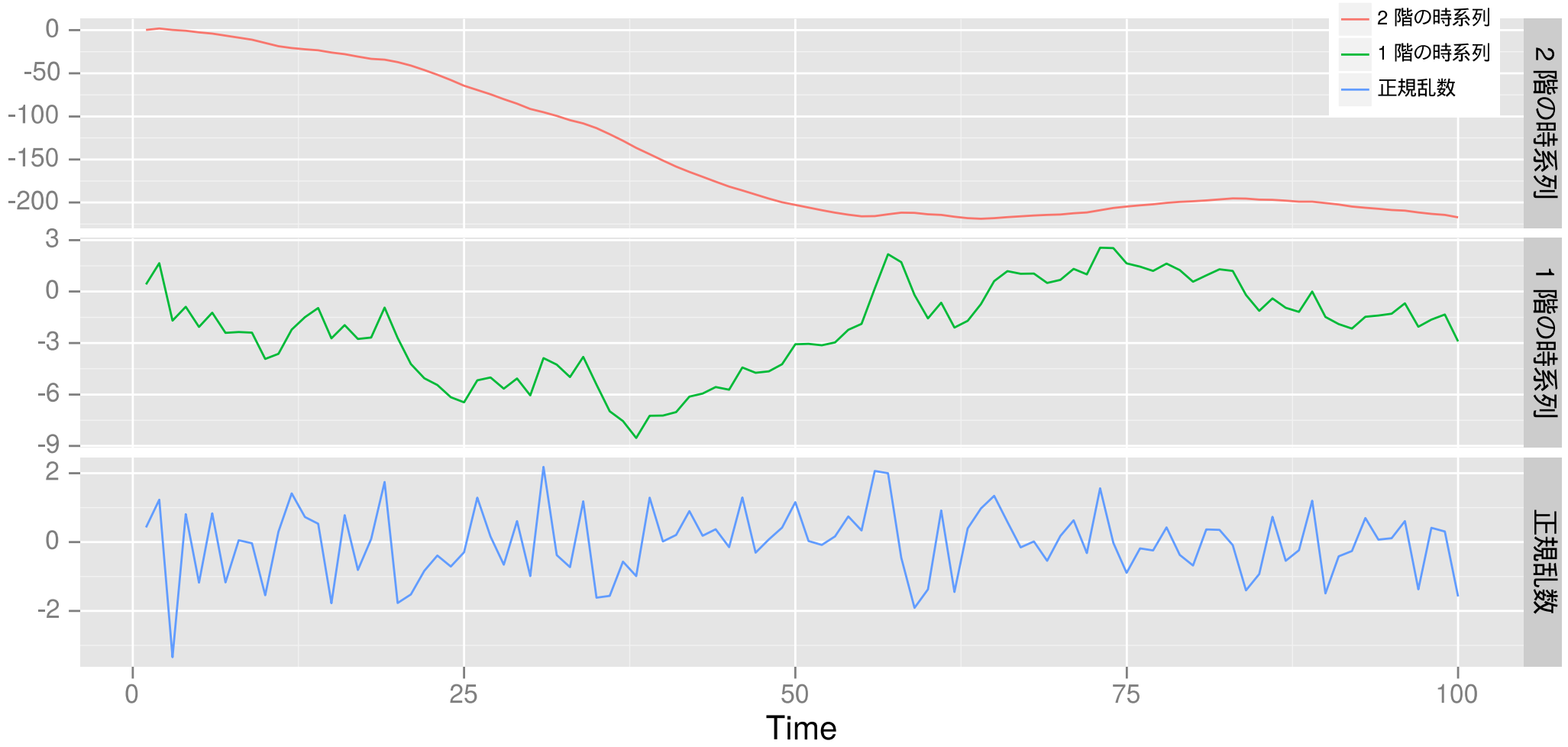
ランダムウォーク

もっとも単純なモデル



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

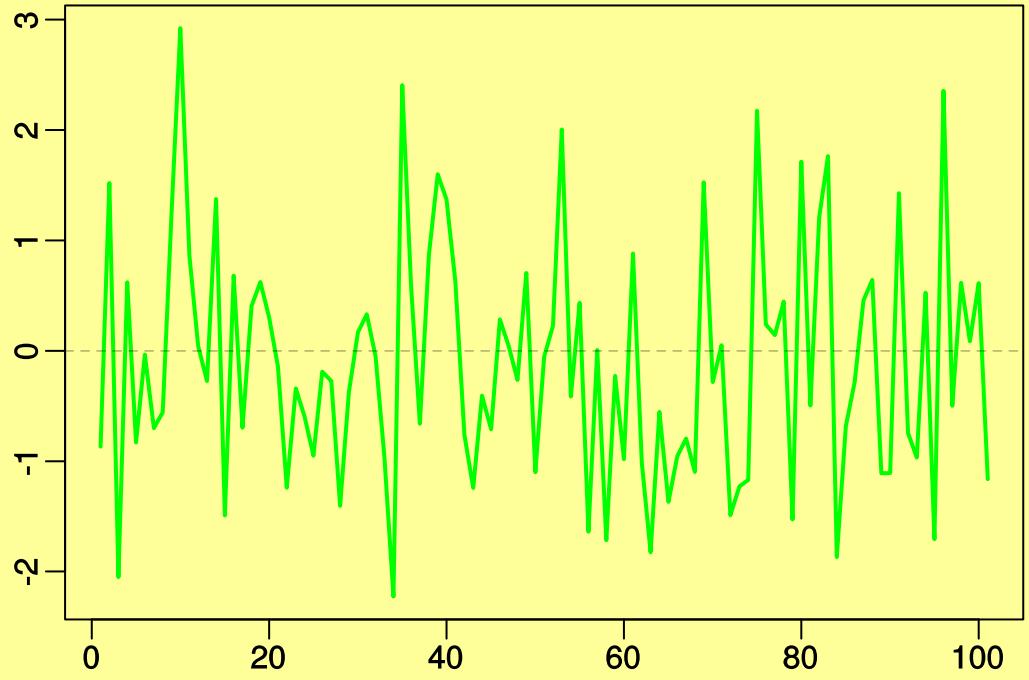


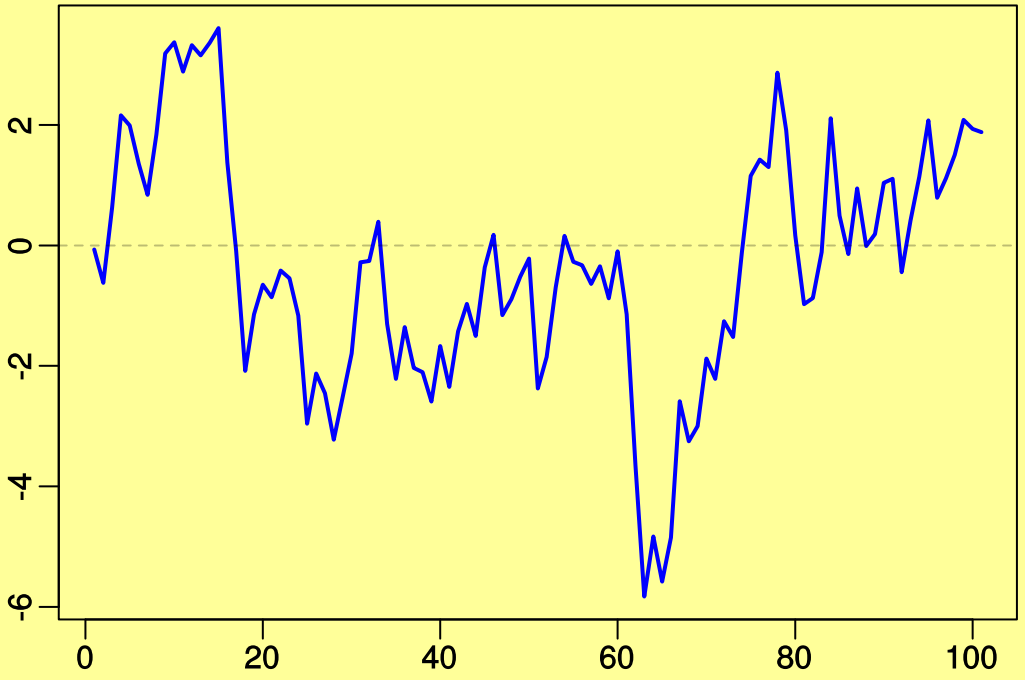
状態空間モデルでたちむかう

時系列データ解析

いろいろな時系列データを

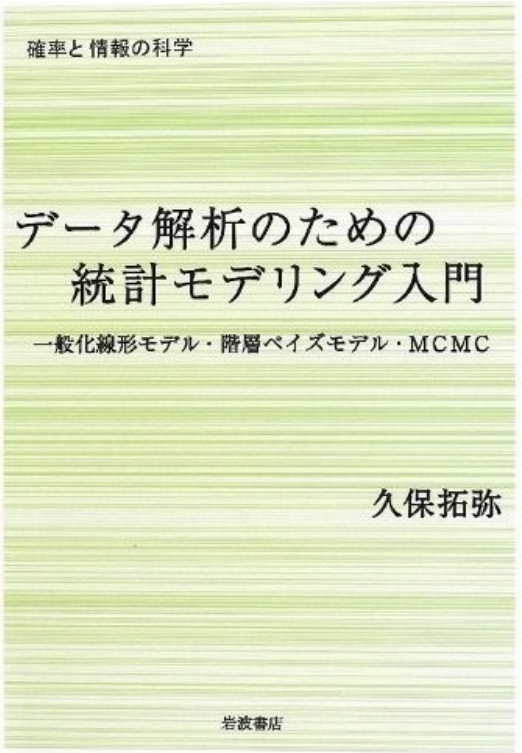
統一的にあつかえないか?





「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!

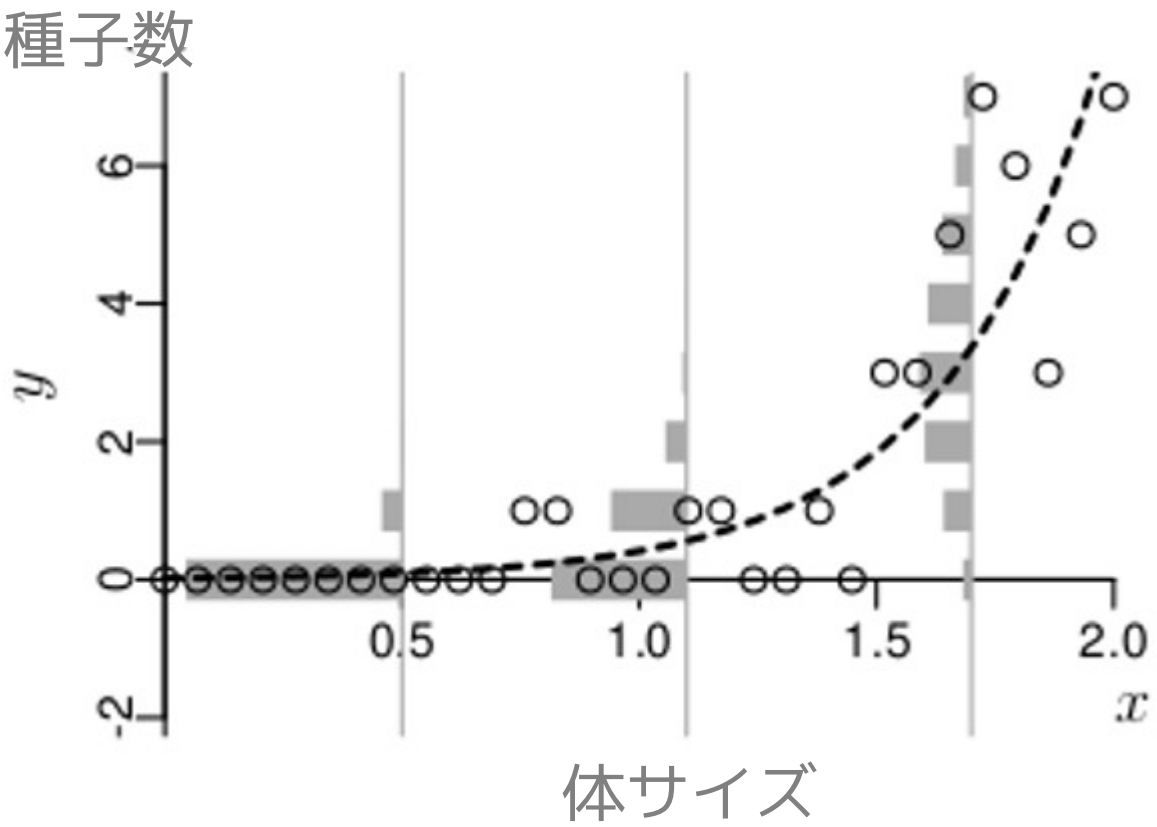
もうすこし「わかった」ような気分?

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?

時系列データ解析の教科書，ねえ……

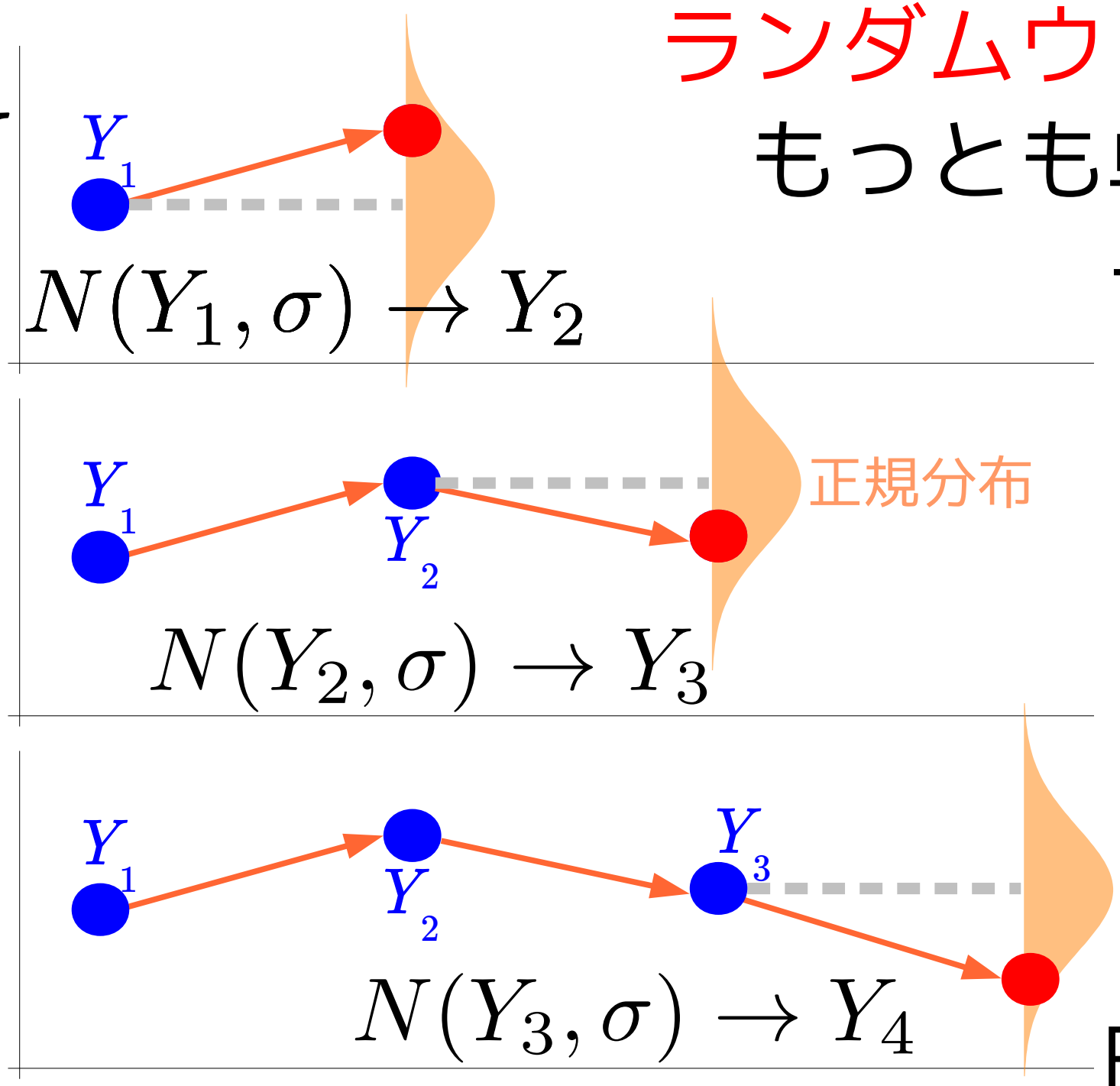
- モデルがあれこれ多すぎる
- 経済学よりのモデルばかり
- なんでも正規分布

なんとかならないかな?

状態空間モデル，どうでしょう?

ランダムウォーク

もっとも単純なモデル



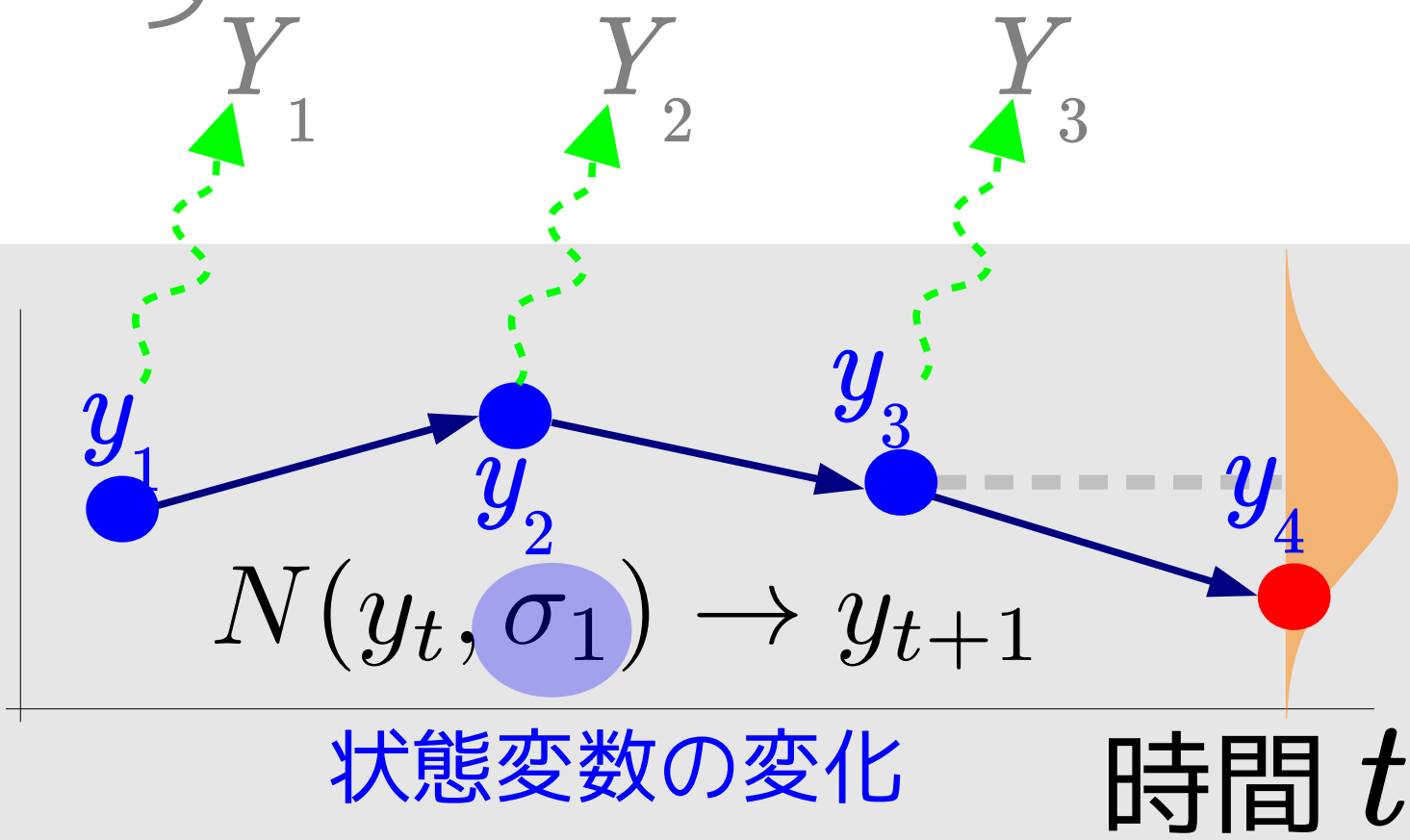
状態空間モデル

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



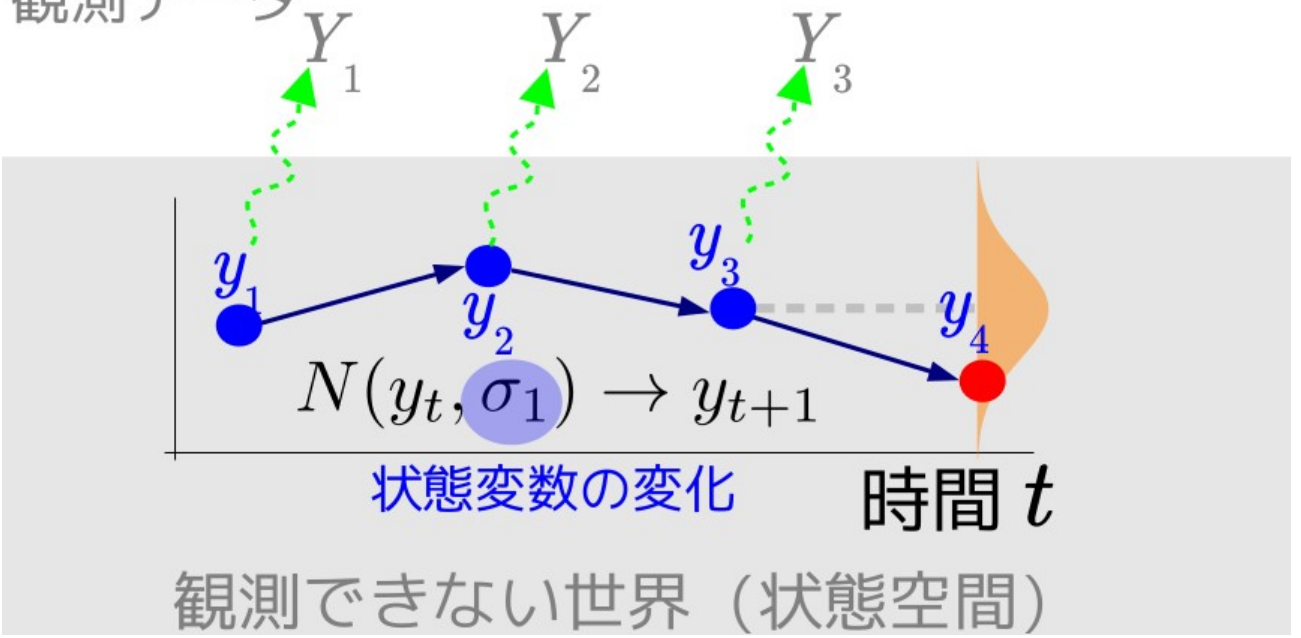
観測できない世界（状態空間）

観測の誤差

状態空間モデル

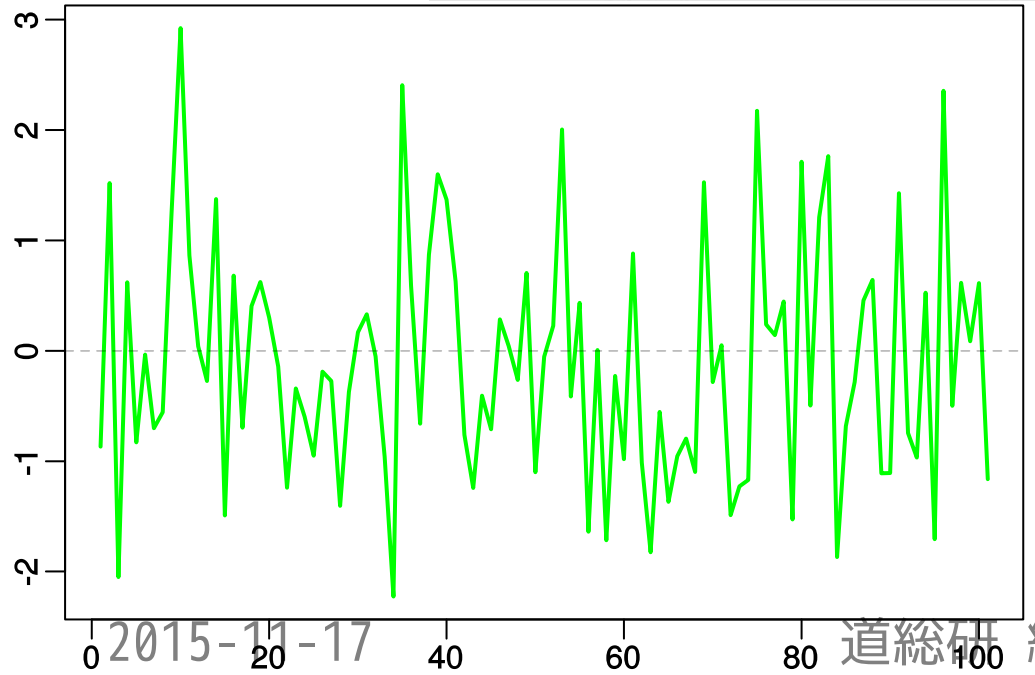
$N(y_t, \sigma_2) \to Y_t$ 二種類の σ をもつ

観測データ

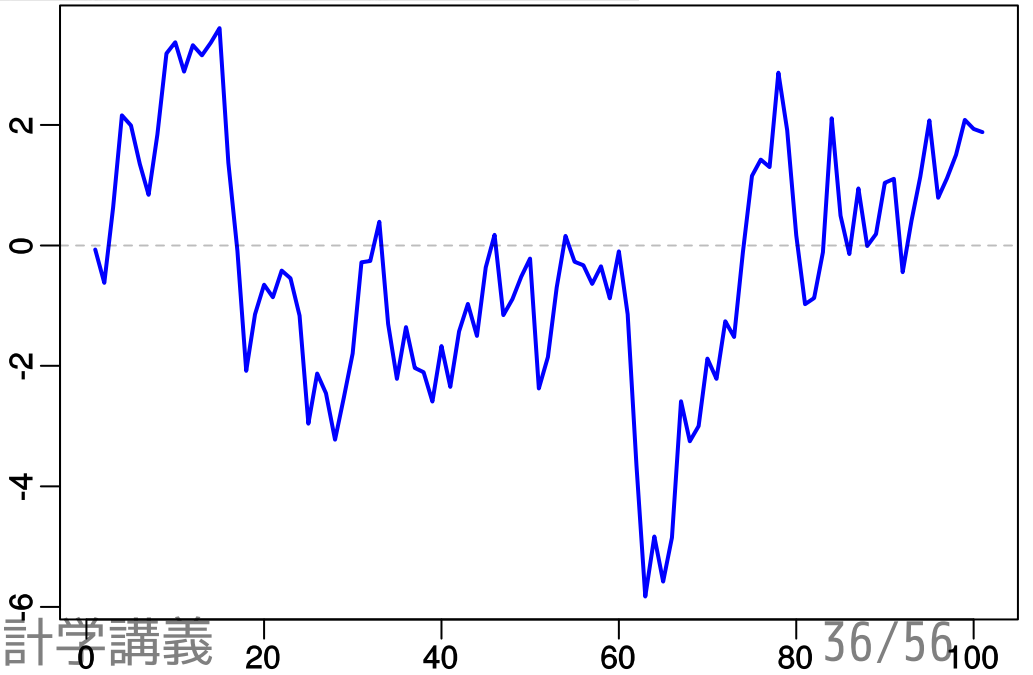


観測できない世界（状態空間）

σ_2 大
σ_1 小



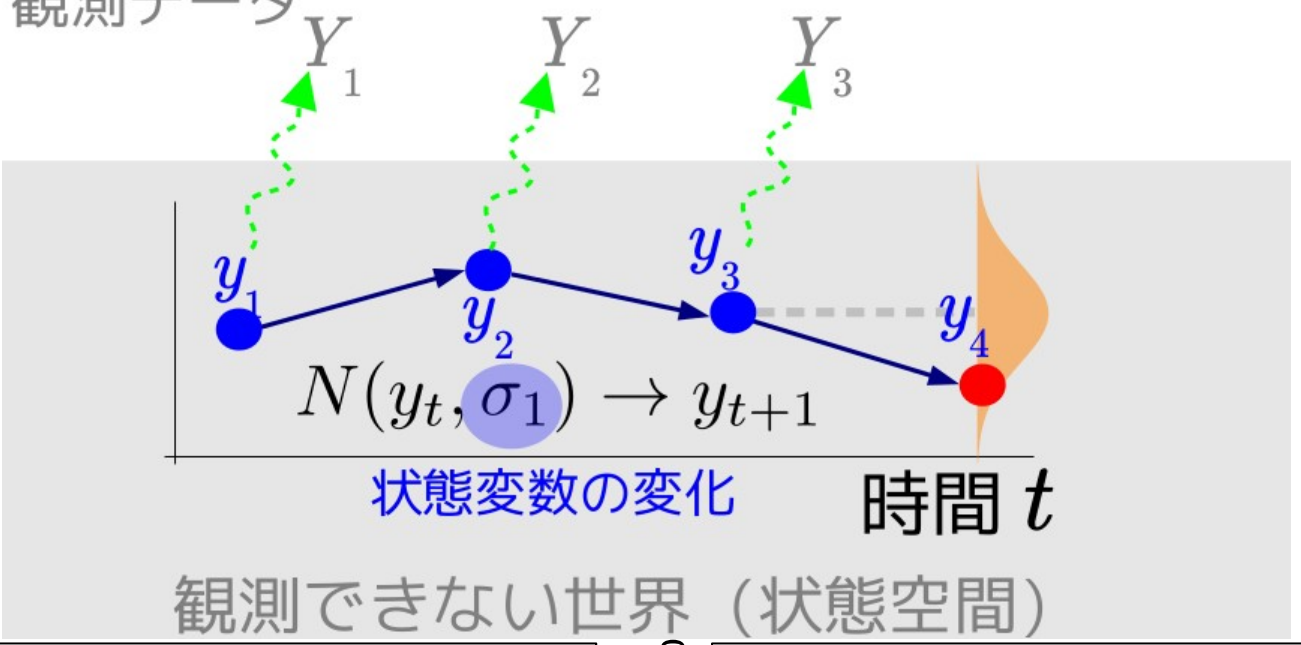
σ_2 小
σ_1 大



観測の誤差　状態空間モデル

$N(y_t, \sigma_2) \to Y_t$　二種類のσをもつ

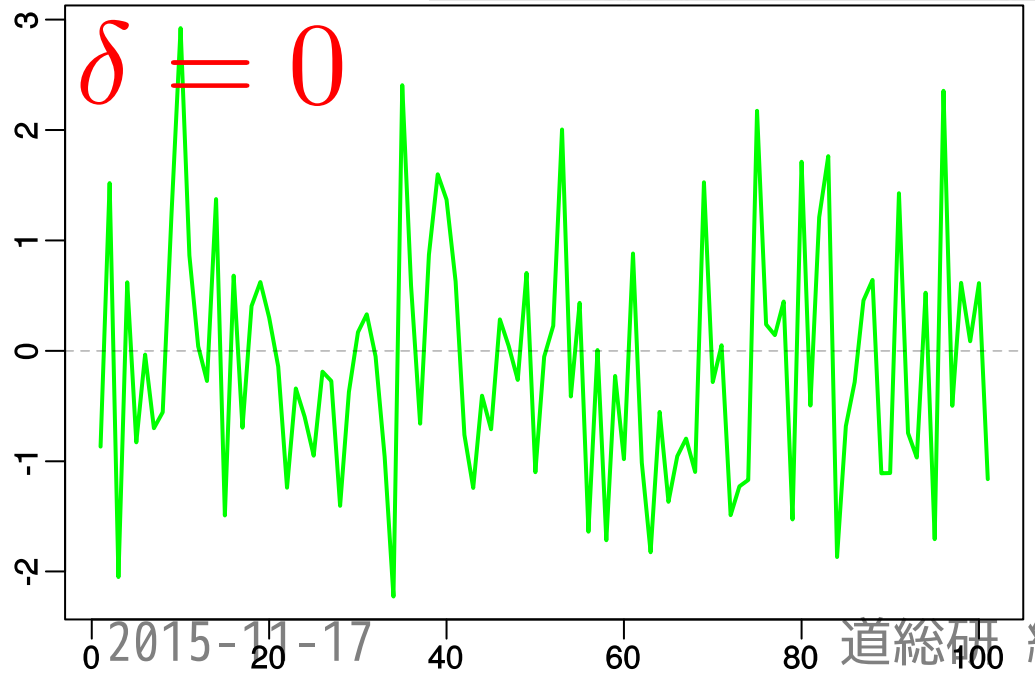
観測データ

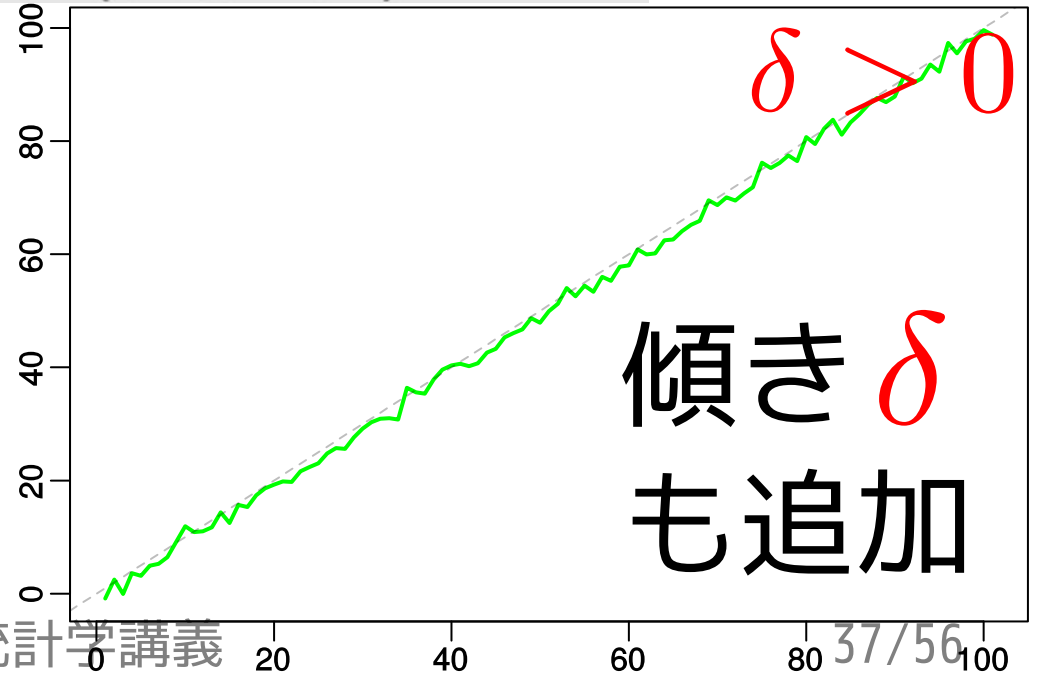


観測できない世界（状態空間）

σ_2 大

σ_1 小



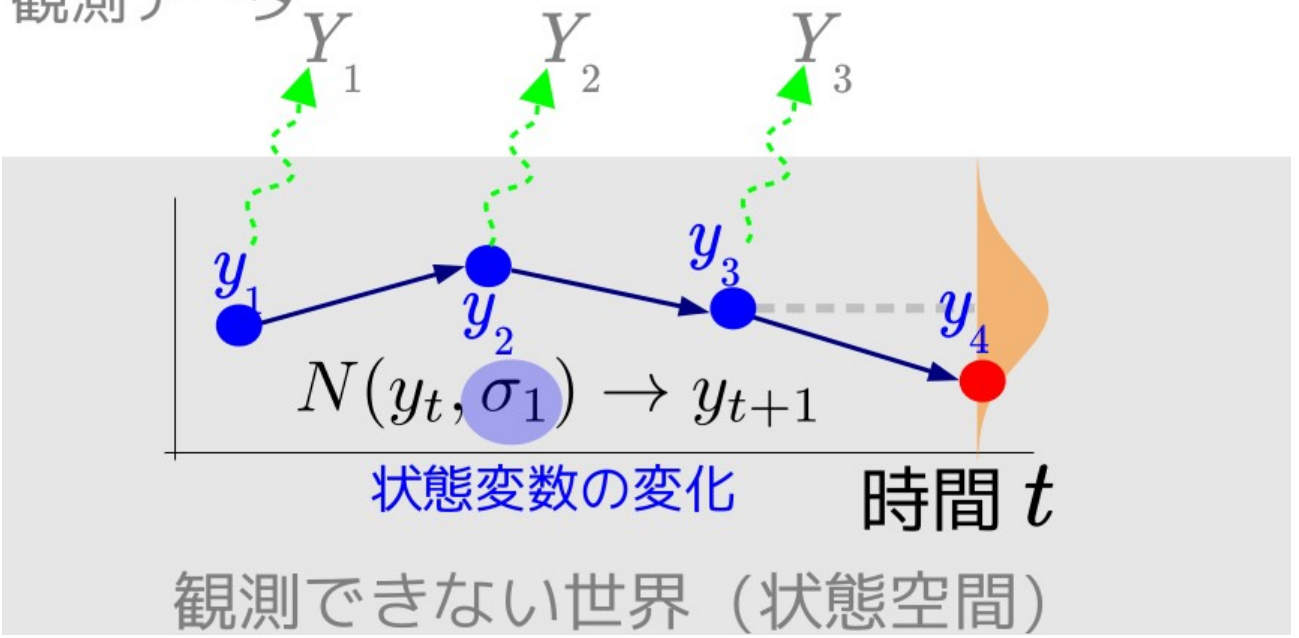


観測の誤差

状態空間モデル

$N(y_t, \sigma_2) \to Y_t$ 二種類の σ をもつ

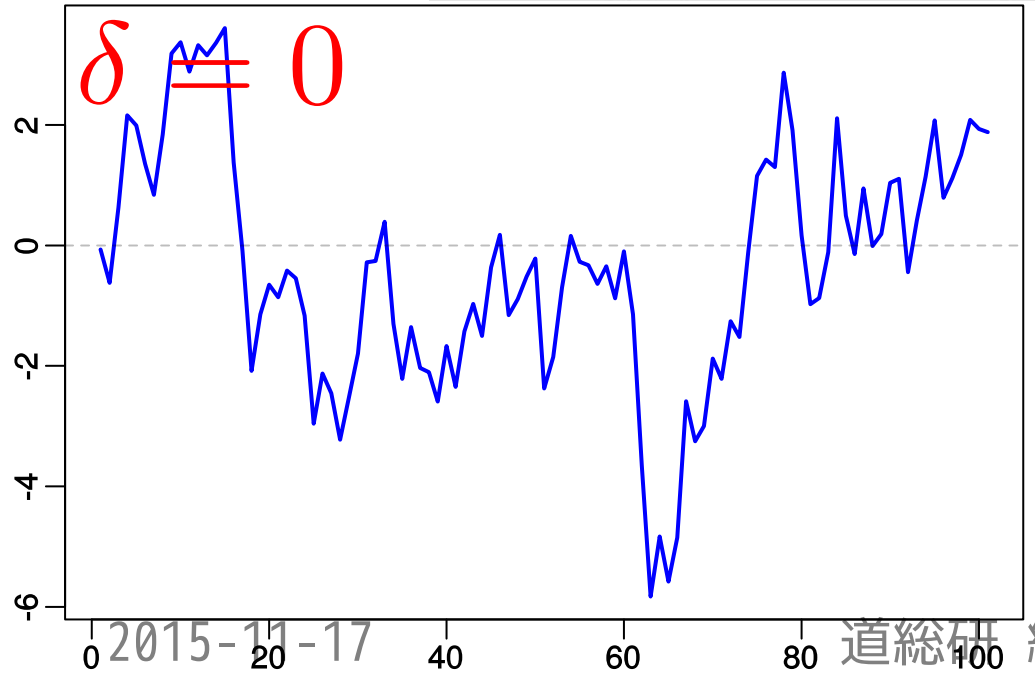
観測データ

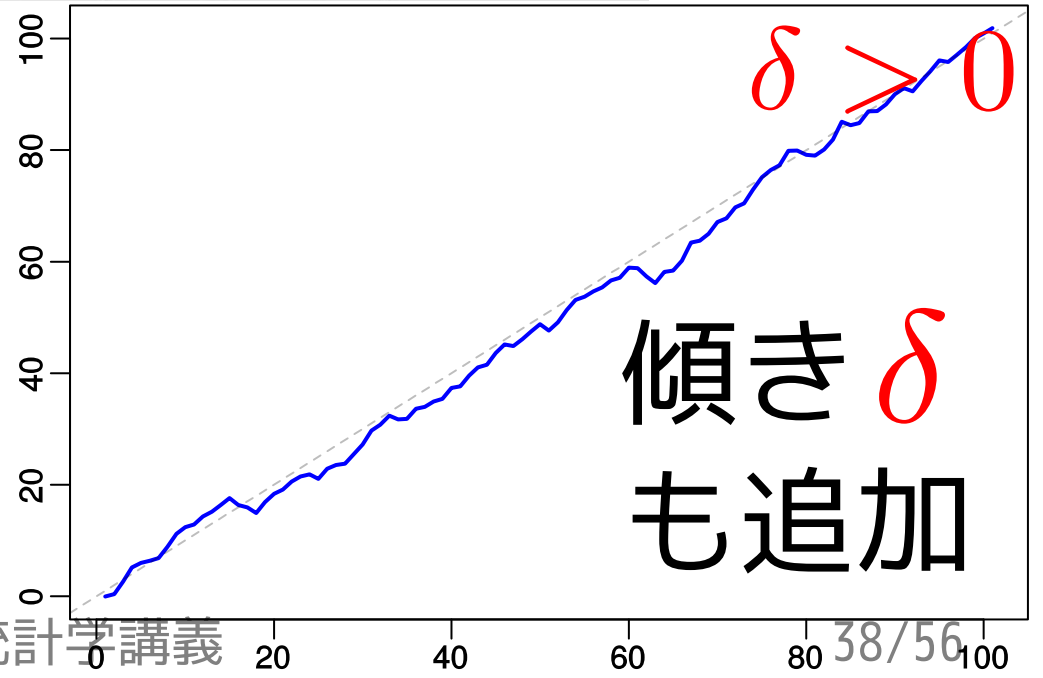


観測できない世界（状態空間）

σ_2 小

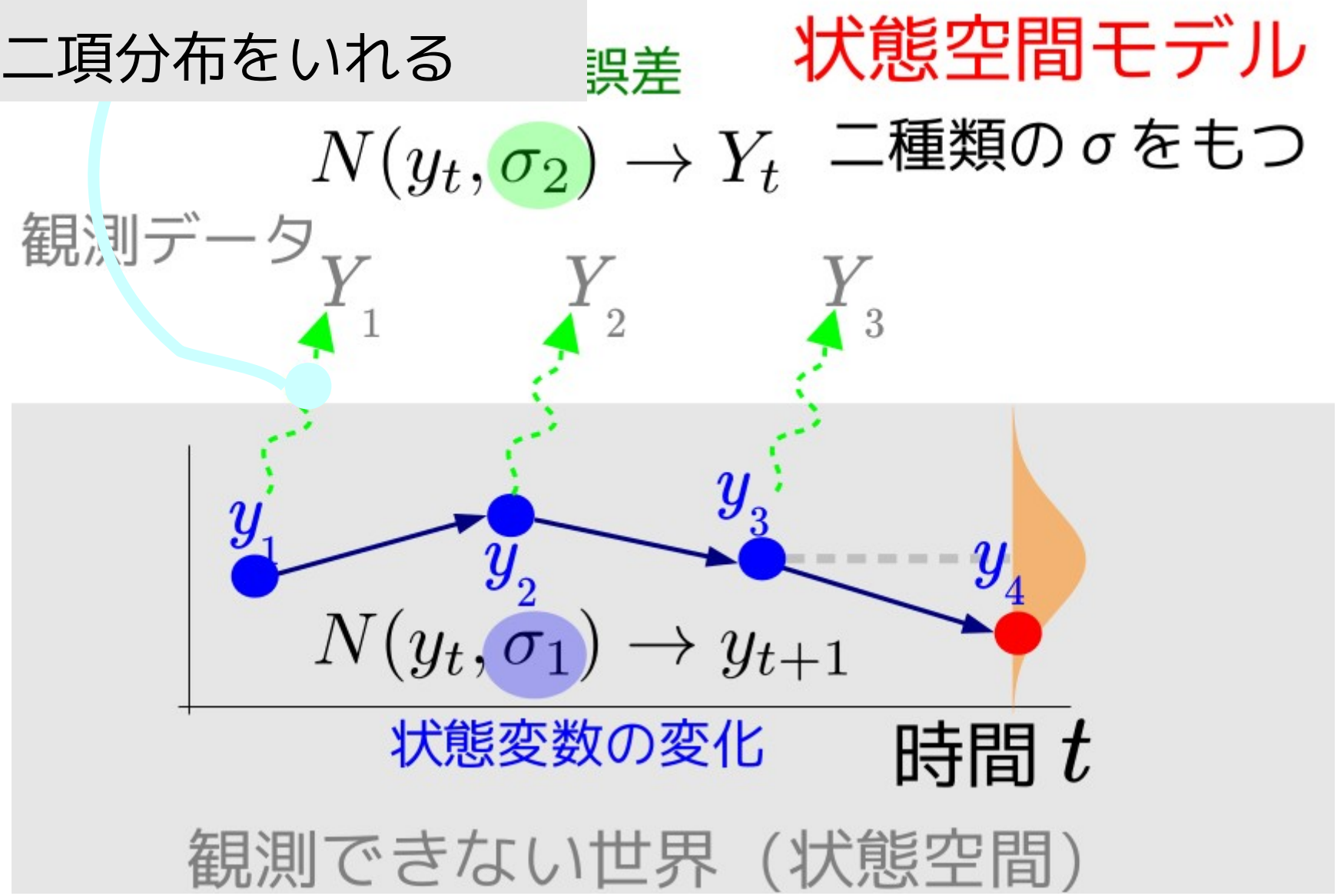
σ_1 大





状態空間モデル + GLM

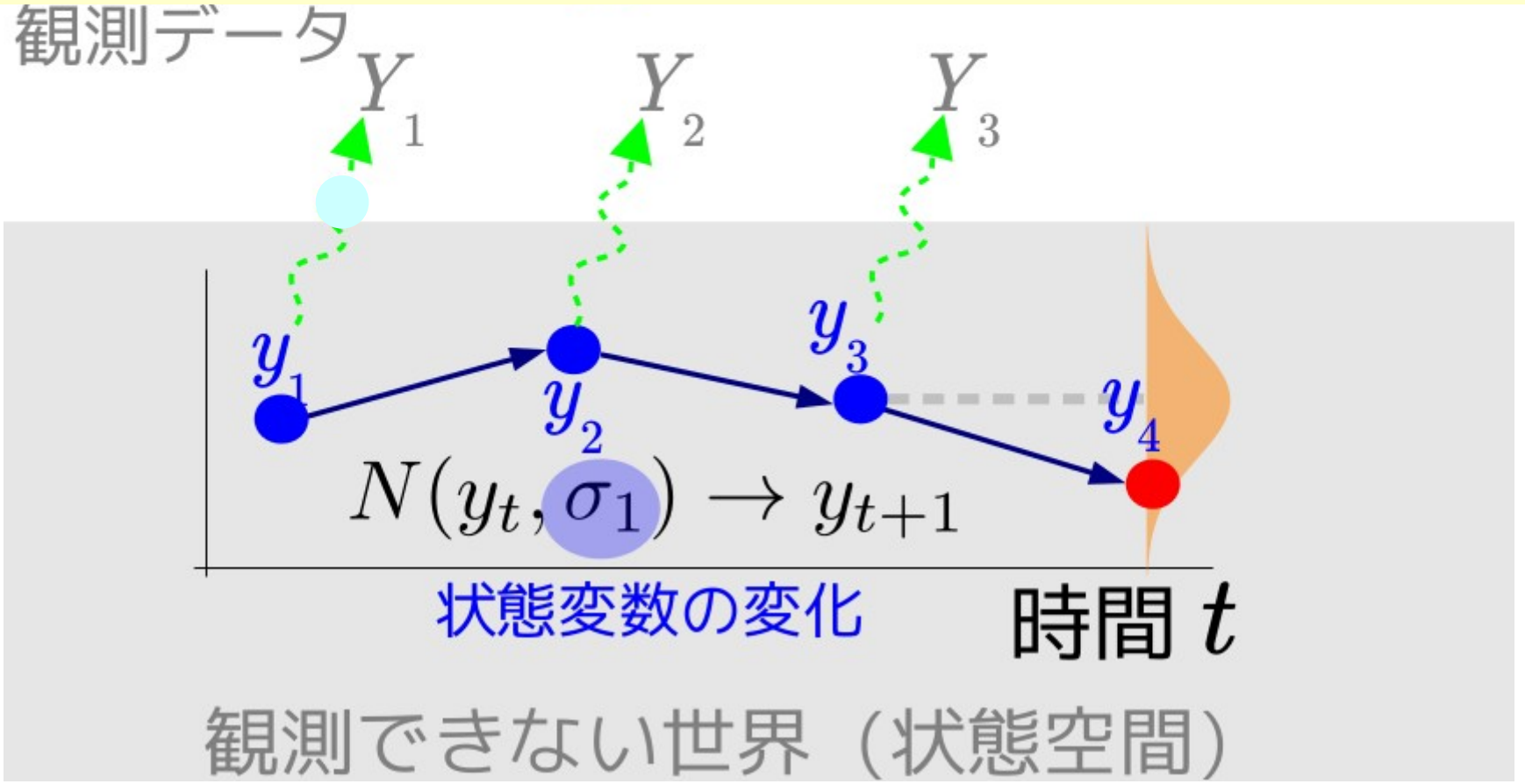
この部分にポアソン分布や
二項分布をいれる



状態空間モデル + GLM

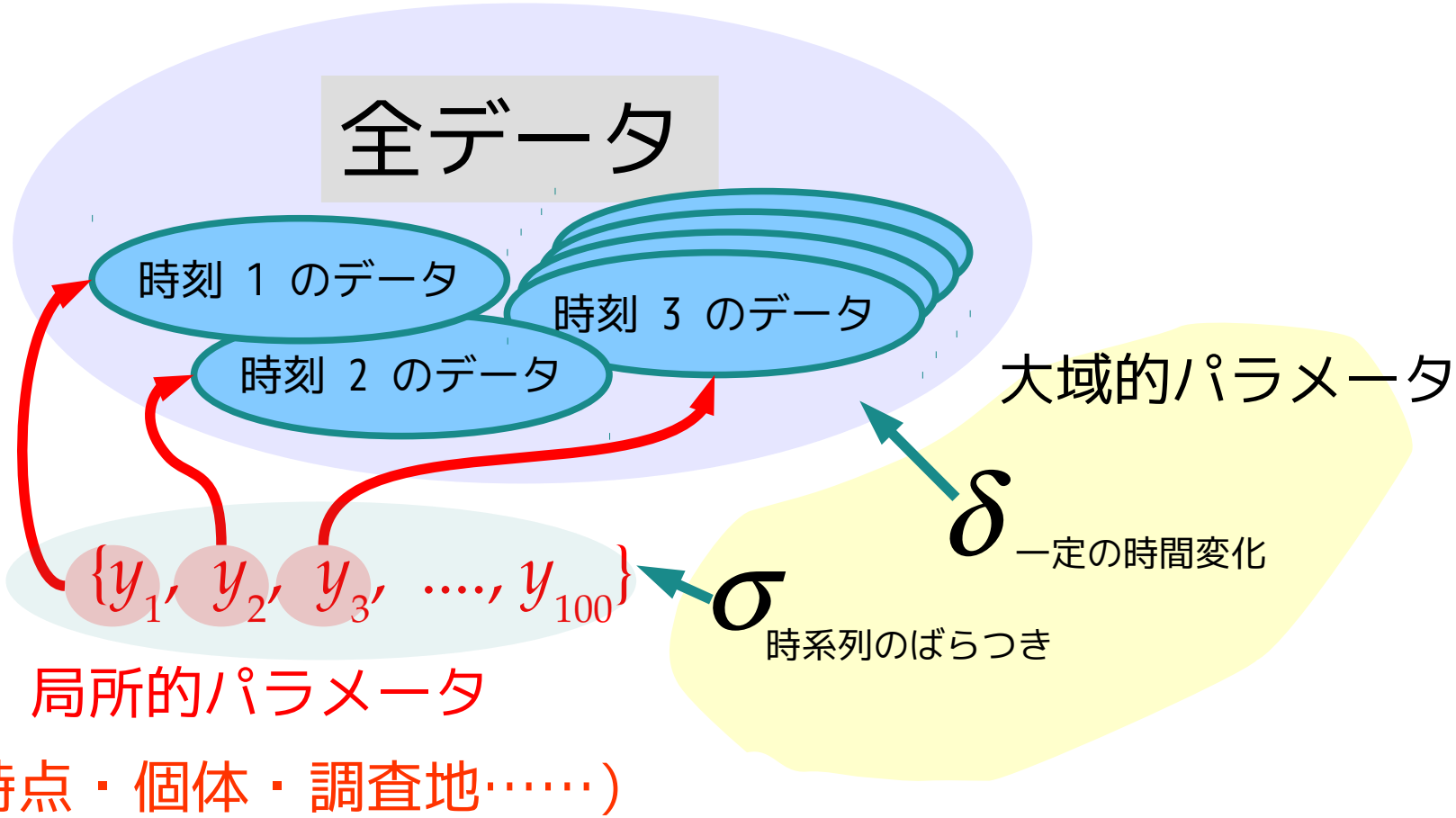
他にも季節変動などを
入れることができます

今日は
省略…
すみません



階層ベイズモデルとは?

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……)

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)

伊東さんが紹介

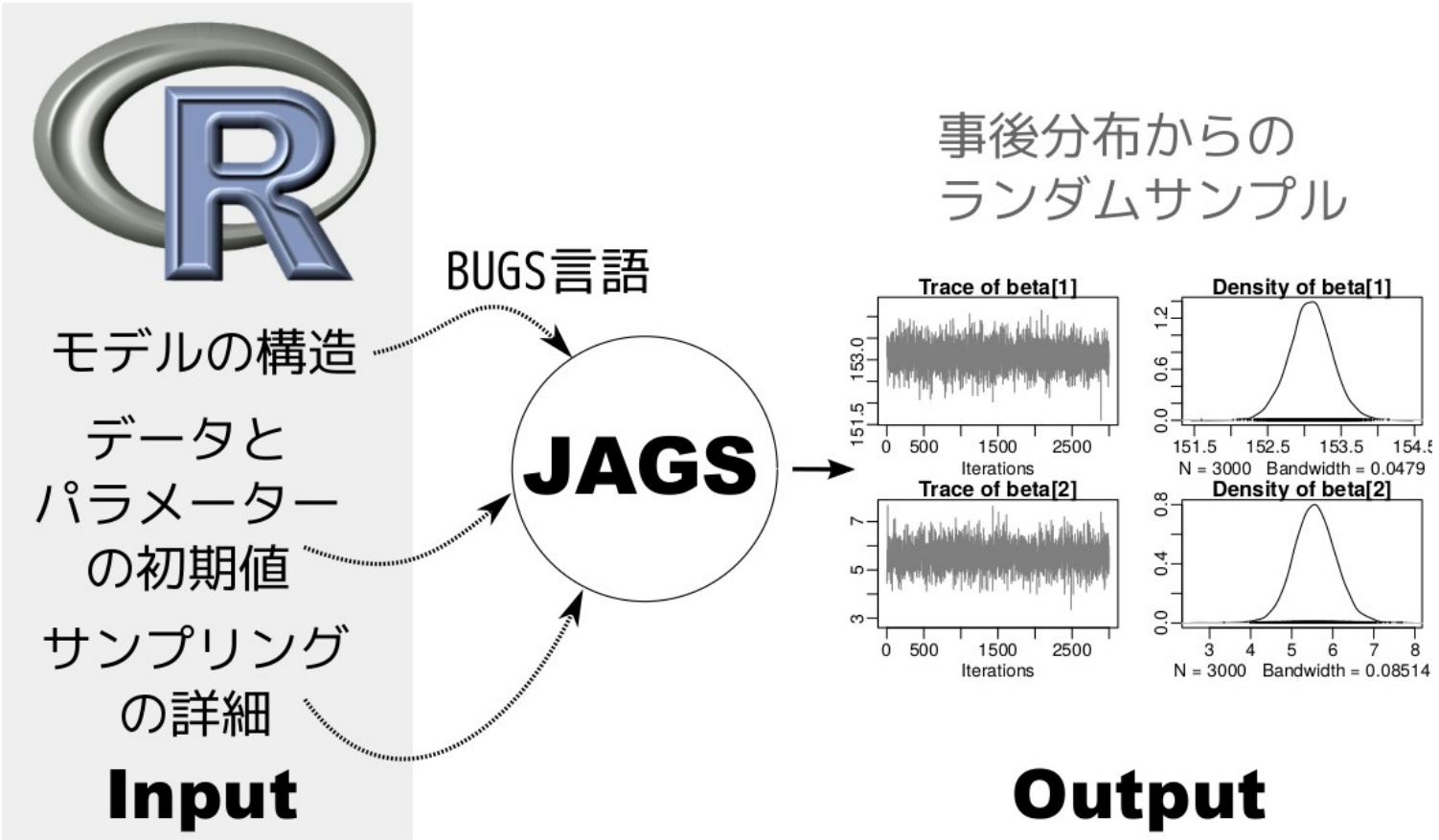
しかしより一般化したモデルに

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

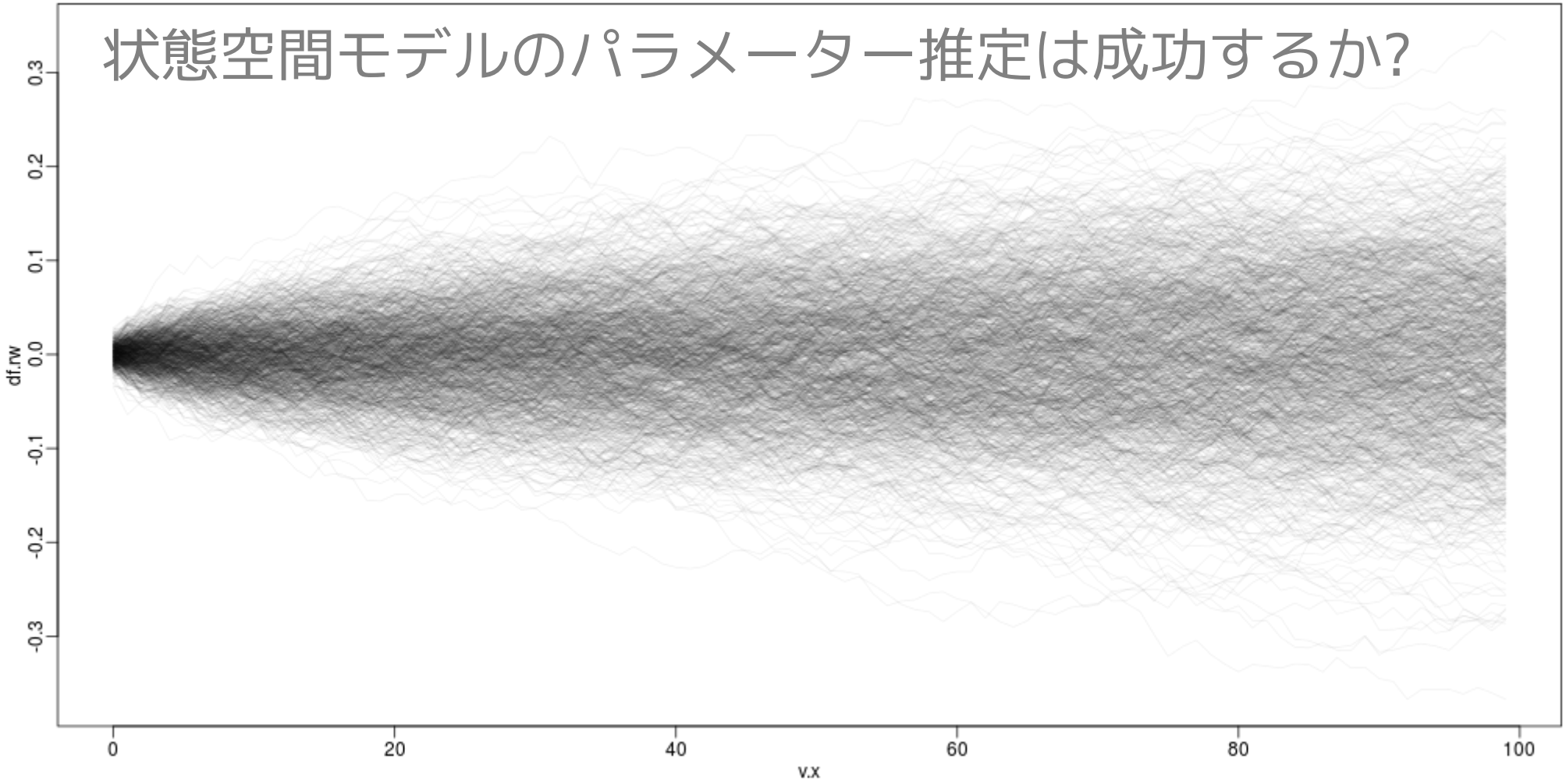
階層ベイズモデルを記述できる



```
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)
    }
}
```

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

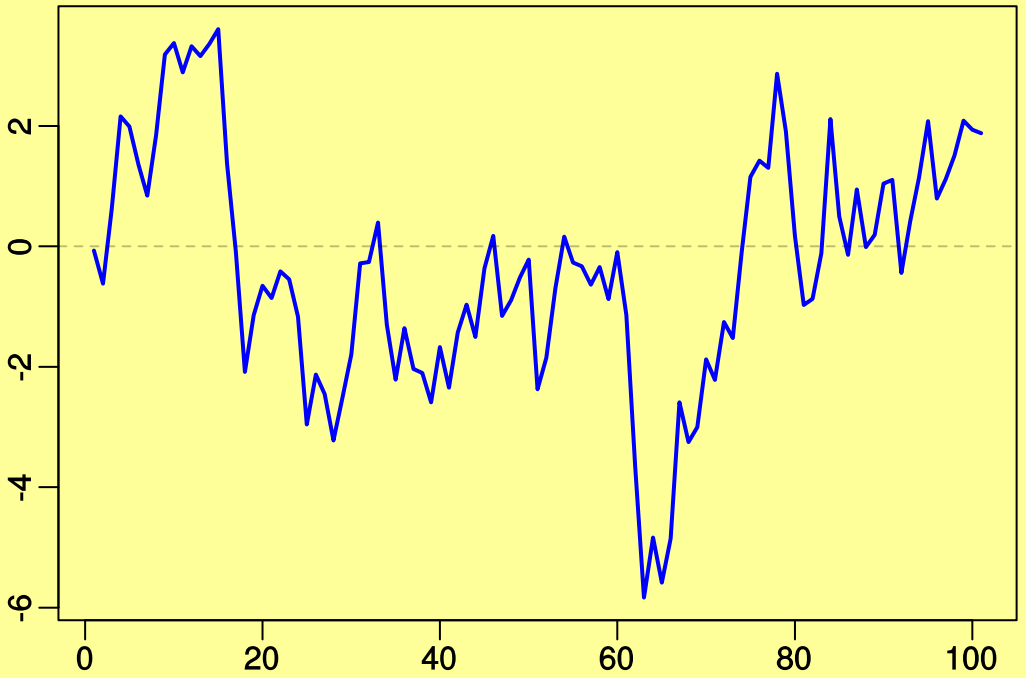
$\delta = 0$

な架空データにあてはめる

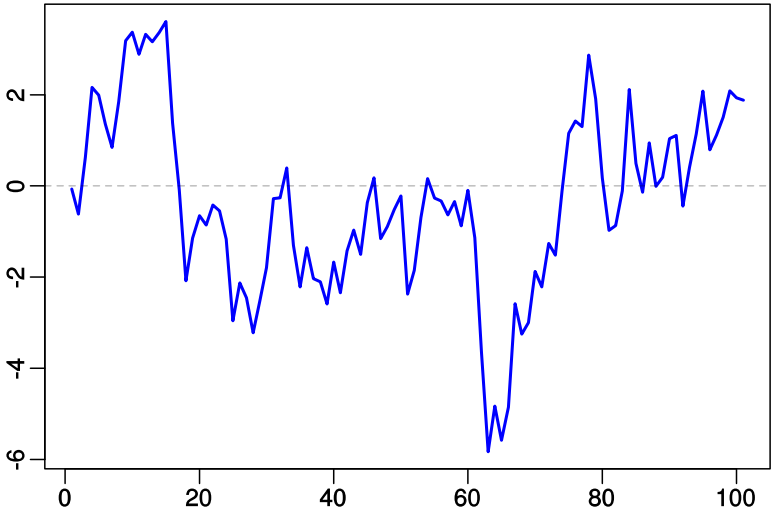
σ_2 小

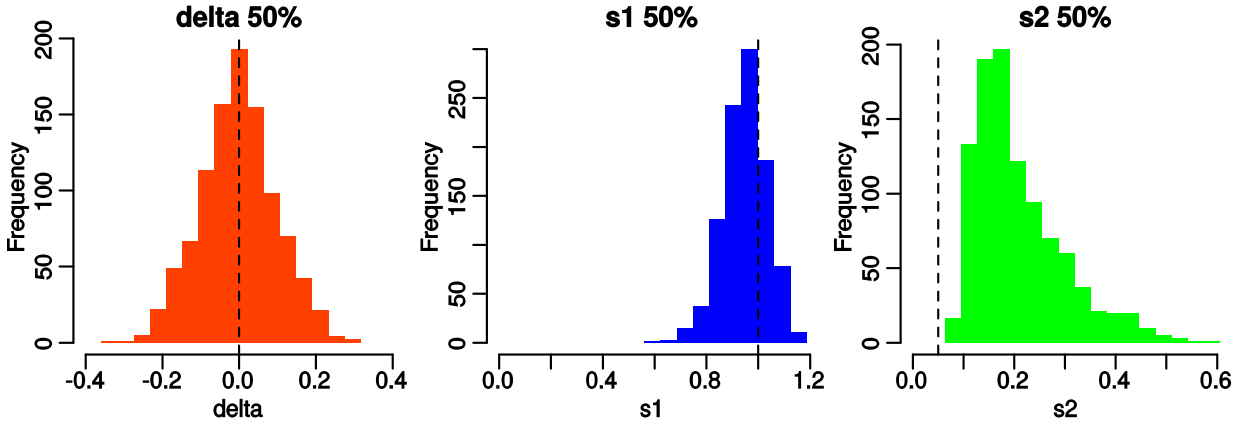
σ_1 大

$\delta = 0$

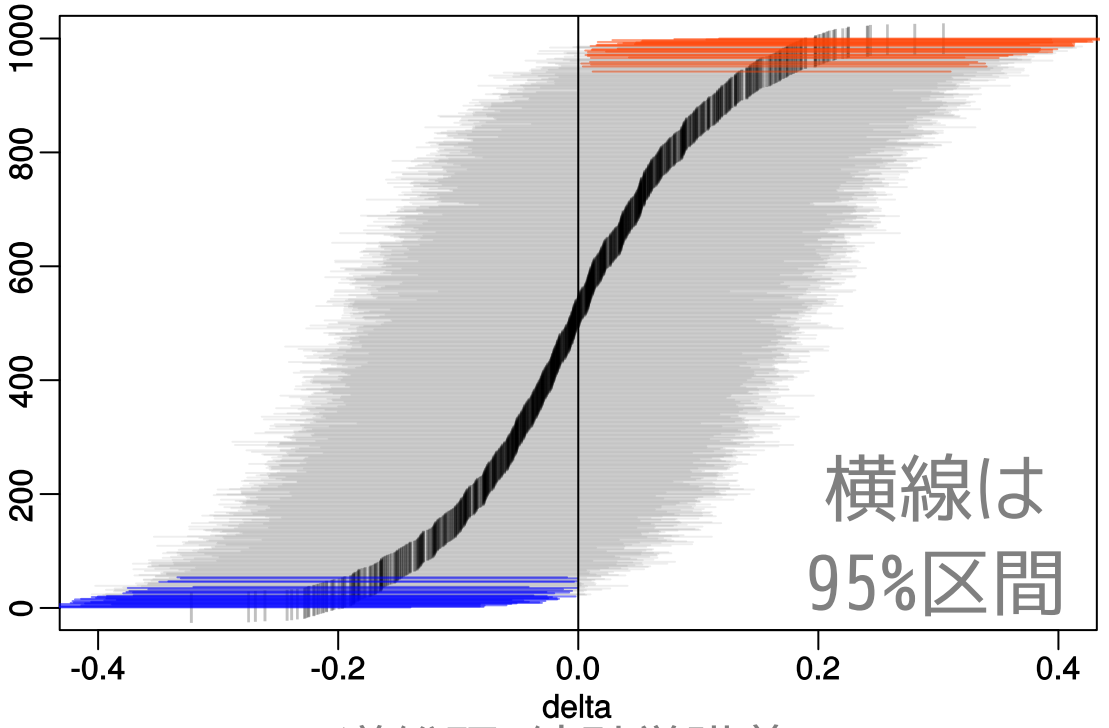


「傾き」δの事後分布を見る





真のδは 0



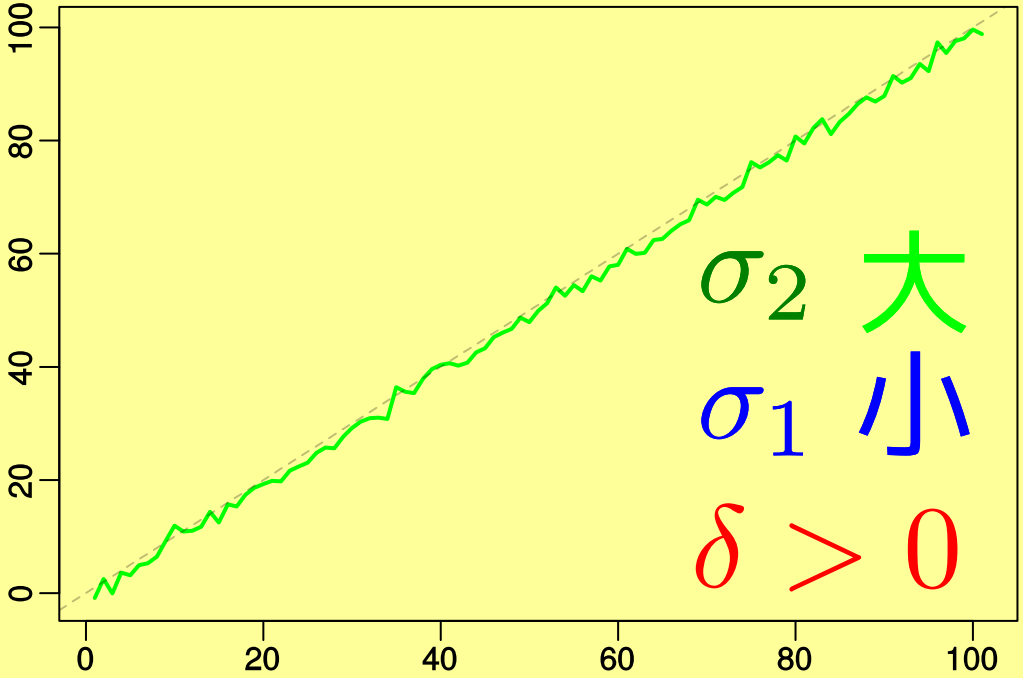
1000回中
63回ずれた

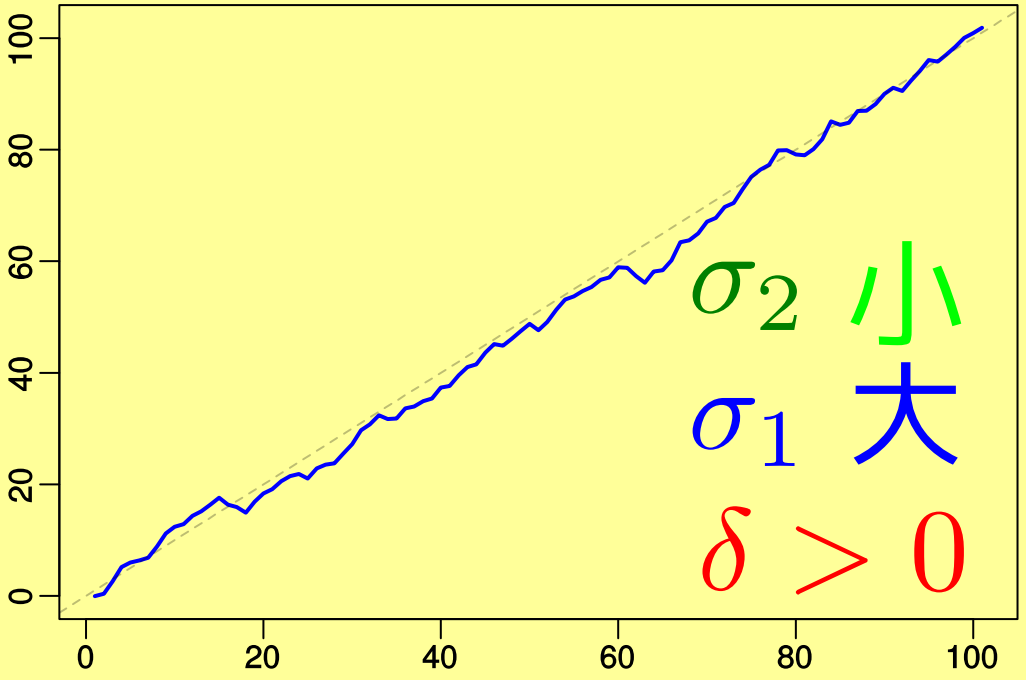
状態空間モデルを

「かたむきあり」ランダムウォーク

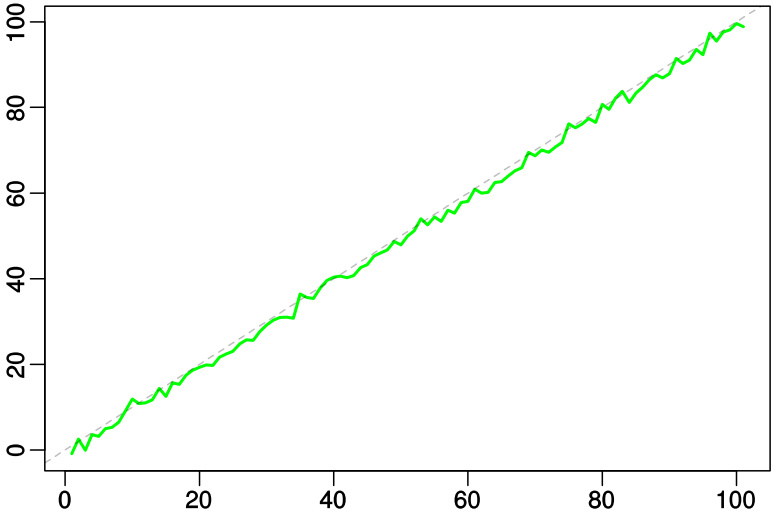
$\delta > 0$

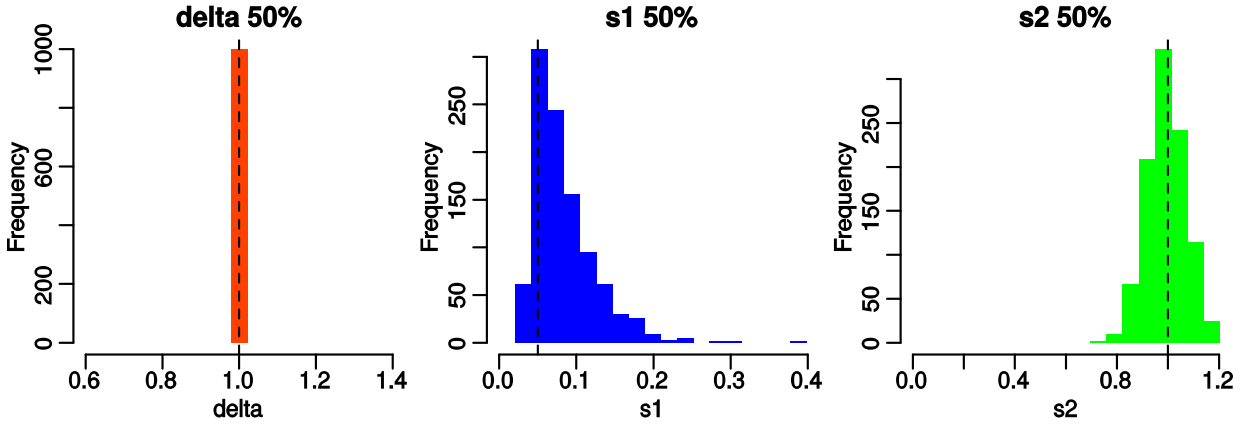
な架空データにあてはめる



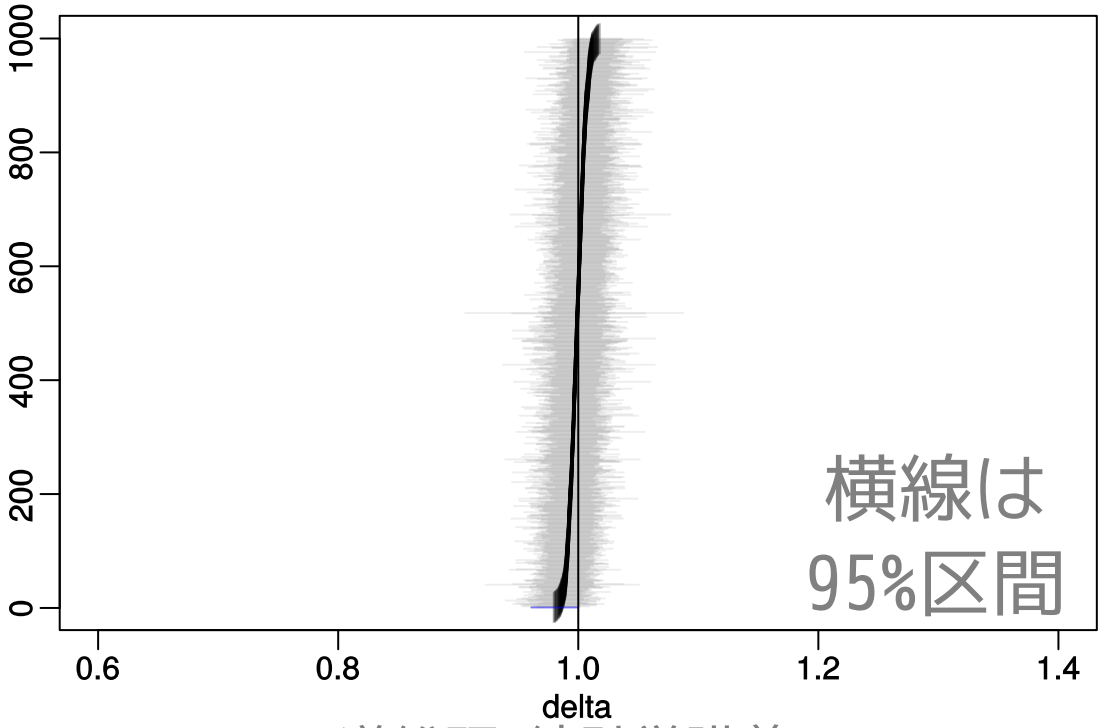


「傾き」δの事後分布を見る



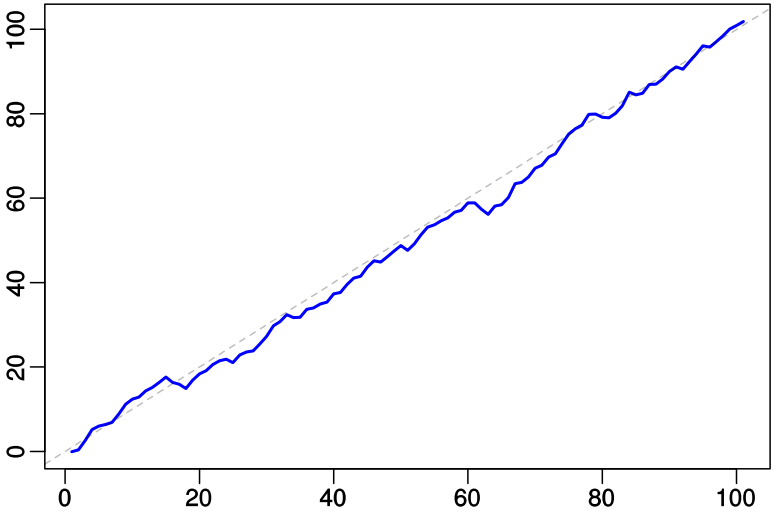


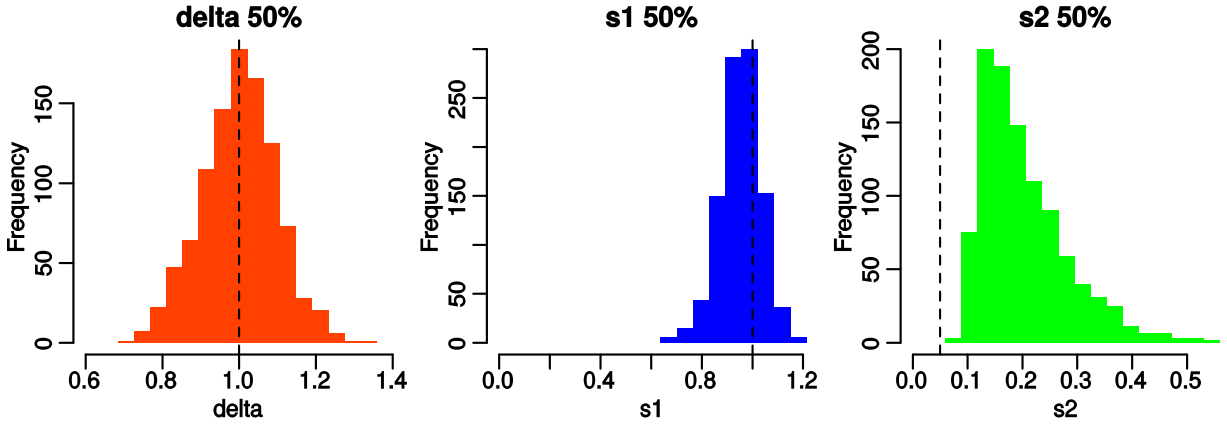
真のδは 1



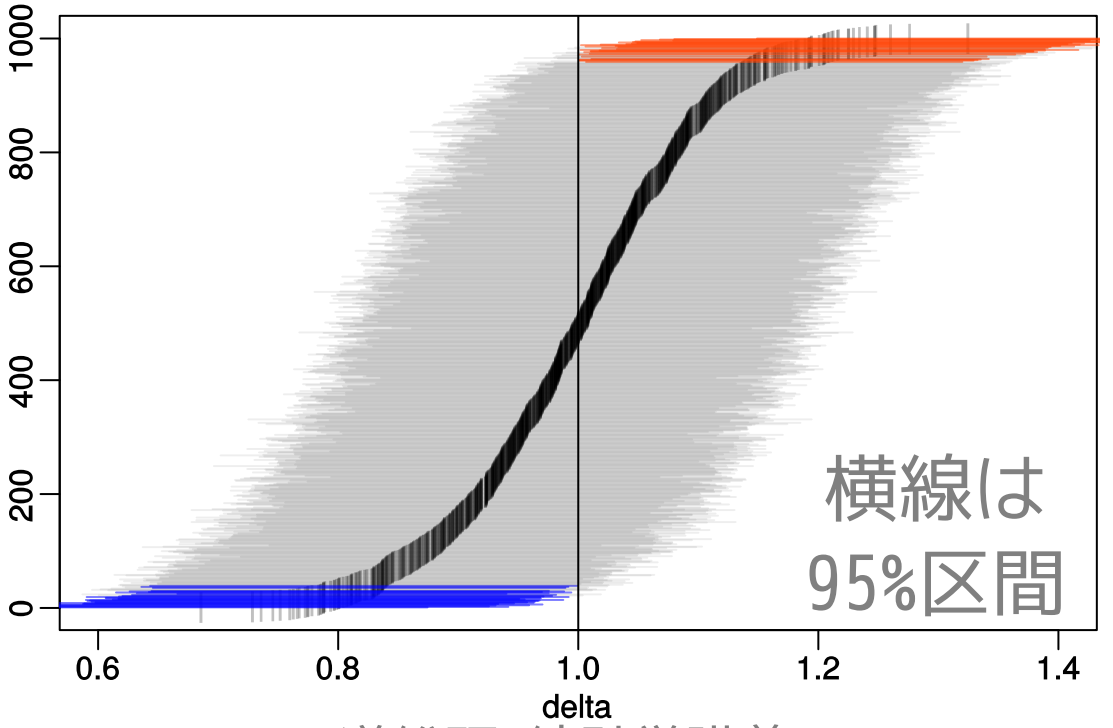
1000回中
1回ずれた

「傾き」δの事後分布を見る



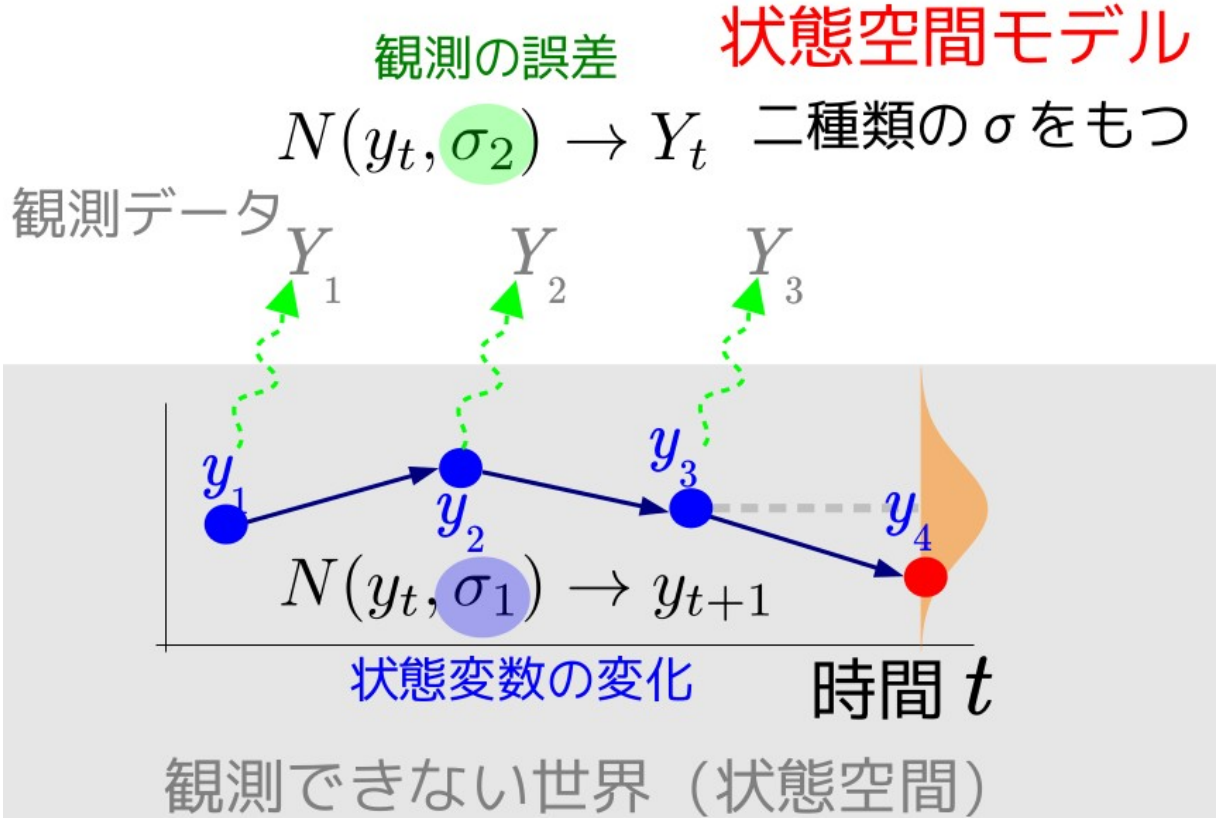


真のδは 1

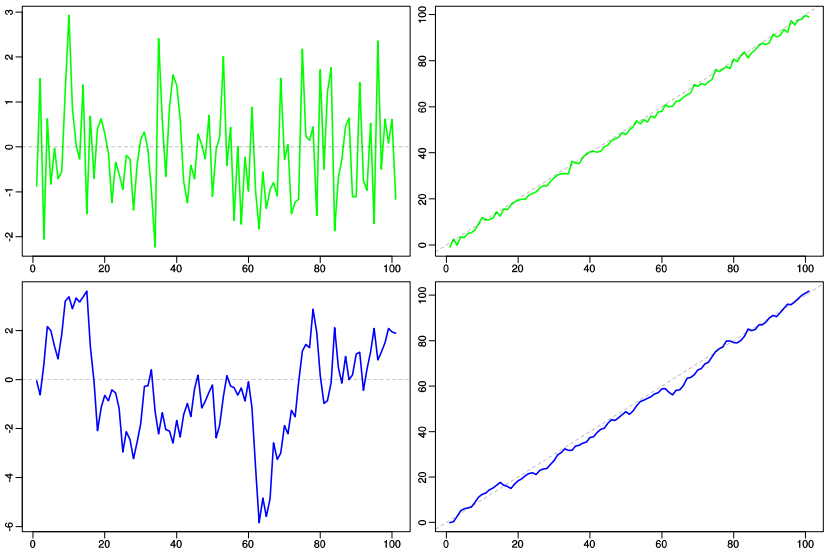


1000回中
62回ずれた

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

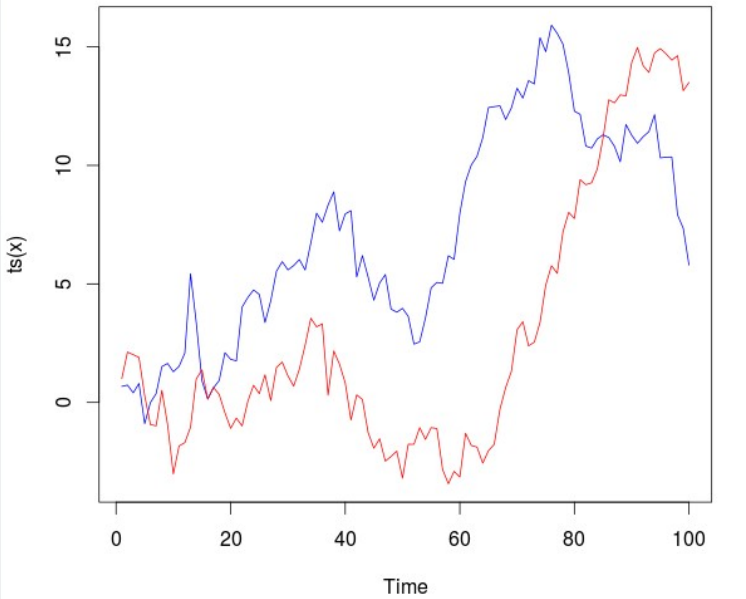
$Y_t \sim X_t$ なうたがわしい回帰

spurious regression

spurious_regression.R

```
1 x <- cumsum(rnorm(100))
2 y <- cumsum(rnorm(100))
3 plot(ts(x), col = "blue", ylim = range(x, y))
4 lines(ts(y), col = "red")
5 print(summary(glm(y ~ x))$coefficients)
```

```
> plot(ts(x), col = "blue", ylim = range(x, y))

> lines(ts(y), col = "red")

> print(summary(glm(y ~ x))$coefficients)
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) -1.67120    0.90288 -1.8510 6.7186e-02
x            0.64551    0.10803  5.9753 3.7127e-08
```



Grenger 因果???

時系列データ解析の

教科書にはよく登場する

複数の時系列感の「相関」

を調べる方法

……

あまり生態学の役には立たないかも

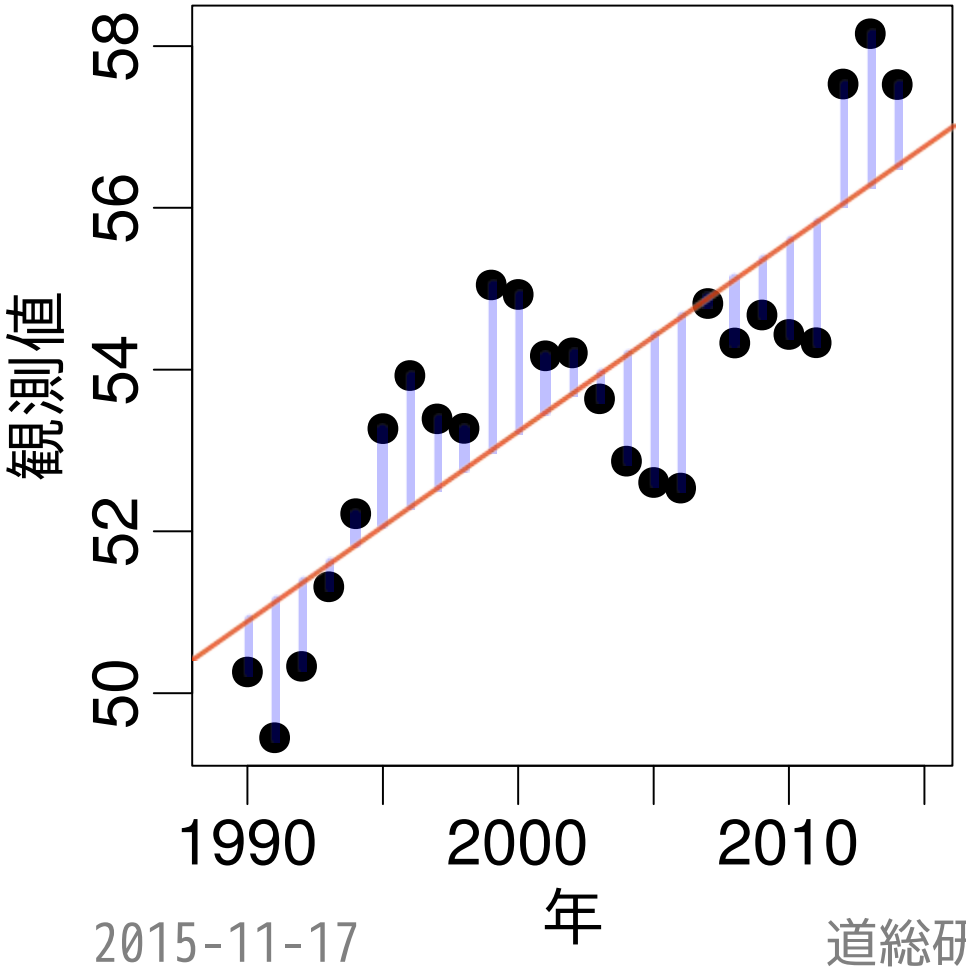
おわりに

時間的な相関はデータの

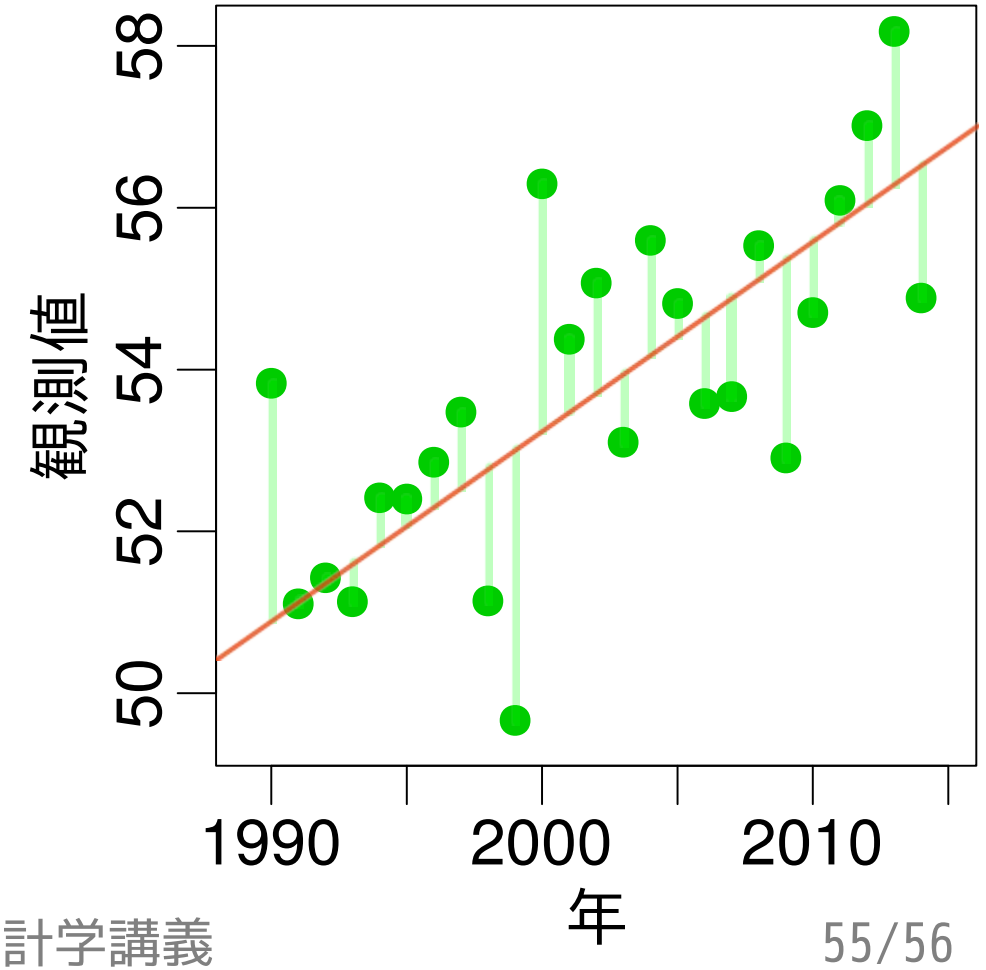
情報量を減少させる

空間相関も…

時系列の「ずれ」



GLM のずれ



時系列データの統計モデリング

- 安易に「回帰」してはいけない
- ランダムウォークモデルが基本
- 統計モデルが生成する時系列パターンを意識する
- 階層ベイズモデルで推定

状態空間モデル