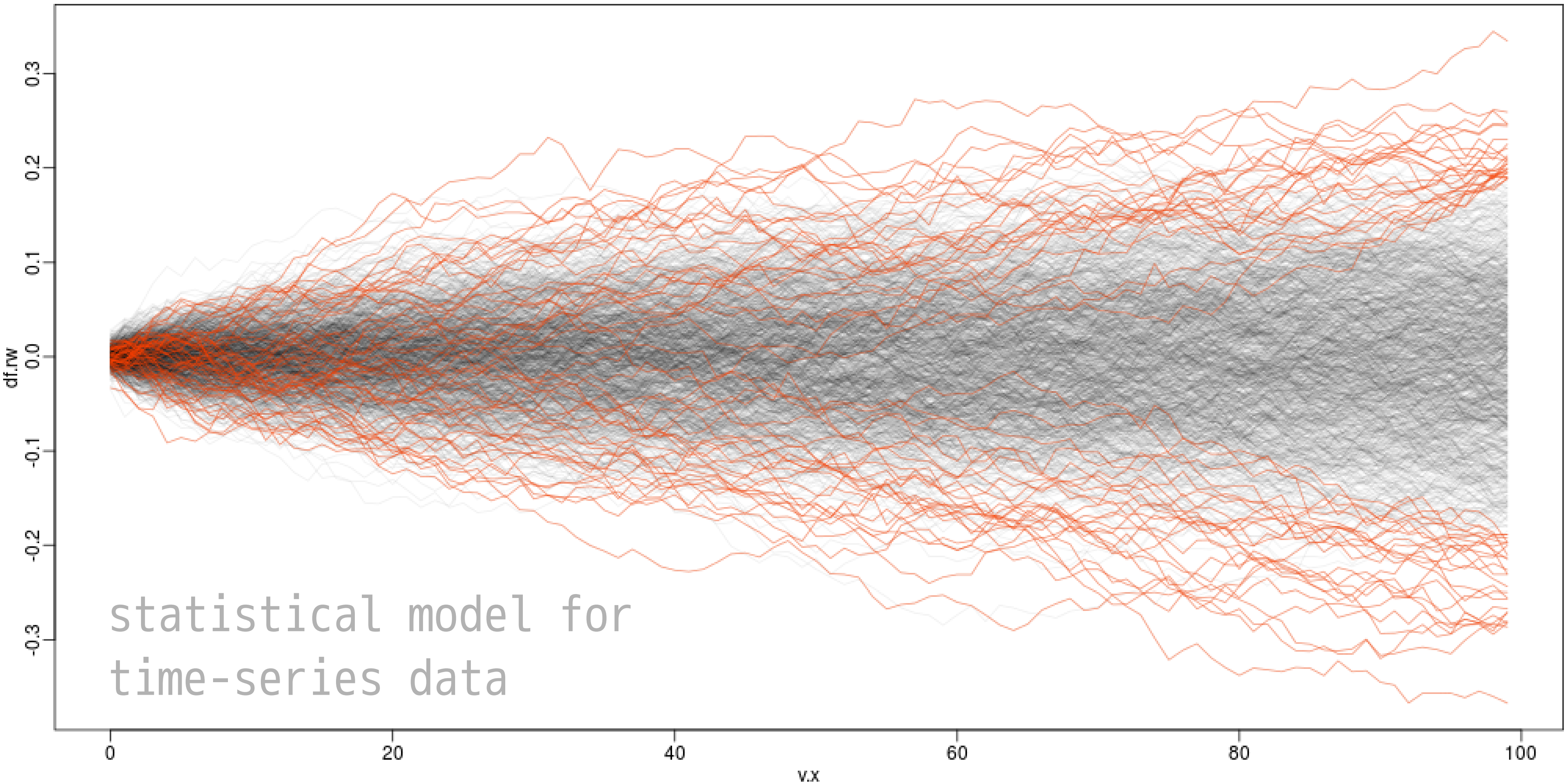
生態学の時系列データ解析でよく見る

『あぶない』モデリング

久保拓弥 mailto:kubo@ees.hokudai.ac.jp



今回・次回の要点

「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

Danger!!

(危1) 時系列データの GLM あてはめ

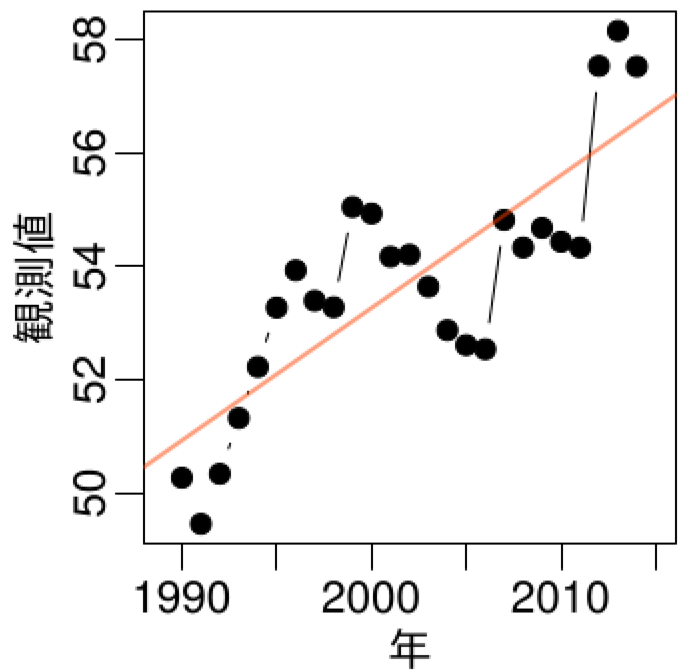
(危2) 時系列 $Y_t \sim$ 時系列 X_t

各時刻の個体数 ～ 気温 とか

(これは次回)

(危 1) 時系列データを GLM で

Do NOT apply GLM to time-series data!



「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?

Danger! time-series Y ~ time-series X

(危2) 時系列 Y_t ～ 時系列 X_t

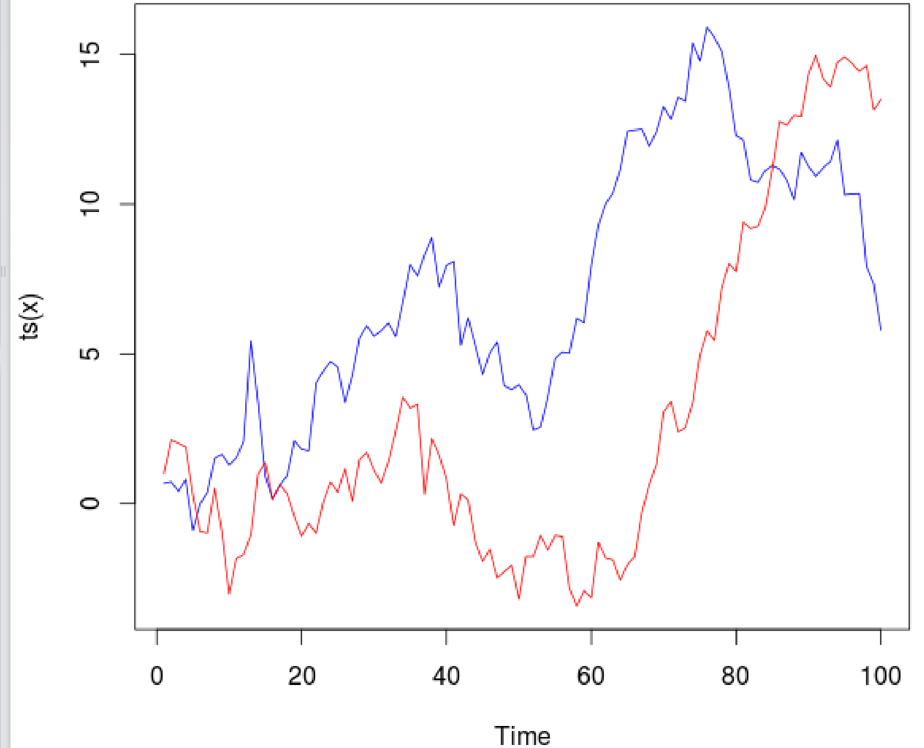
「見せかけの回帰」spurious regression

spurious_regression.R

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

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
>
```



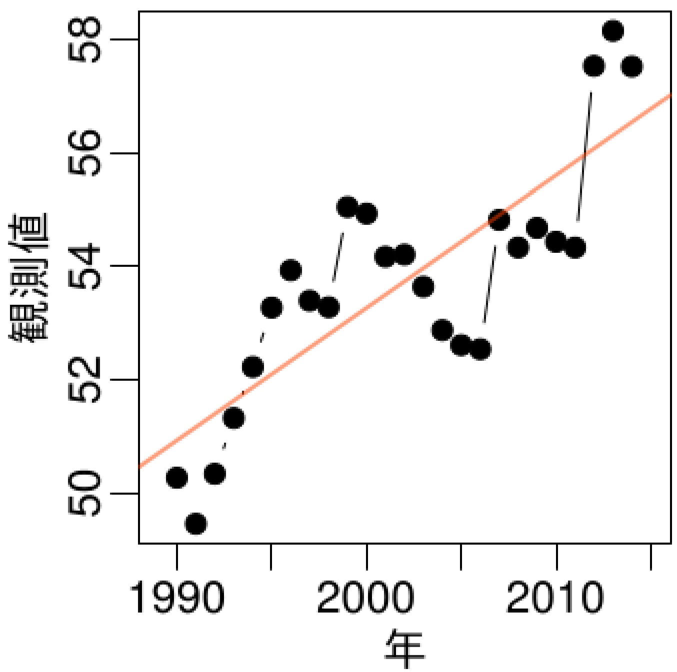
No! Time_series y ~ Time_series x

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

Use state-space models 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

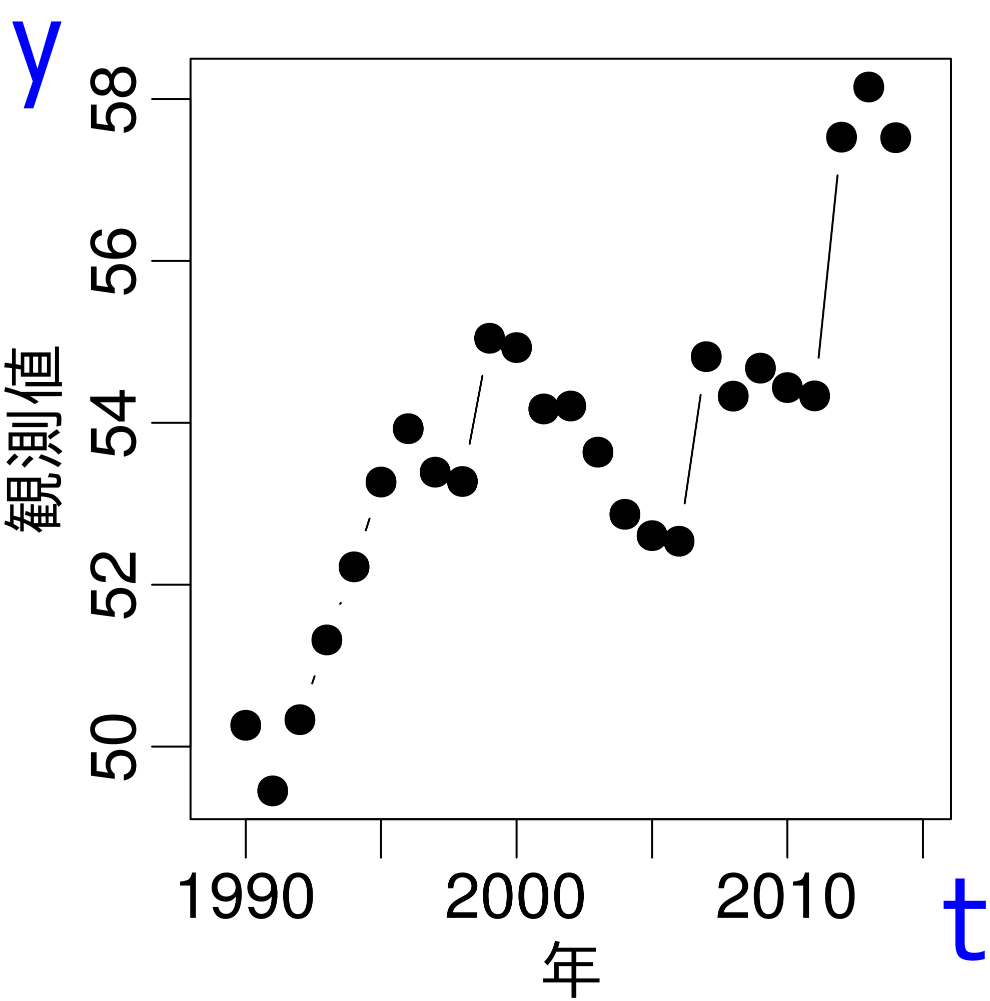
AIC モデル選択

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?

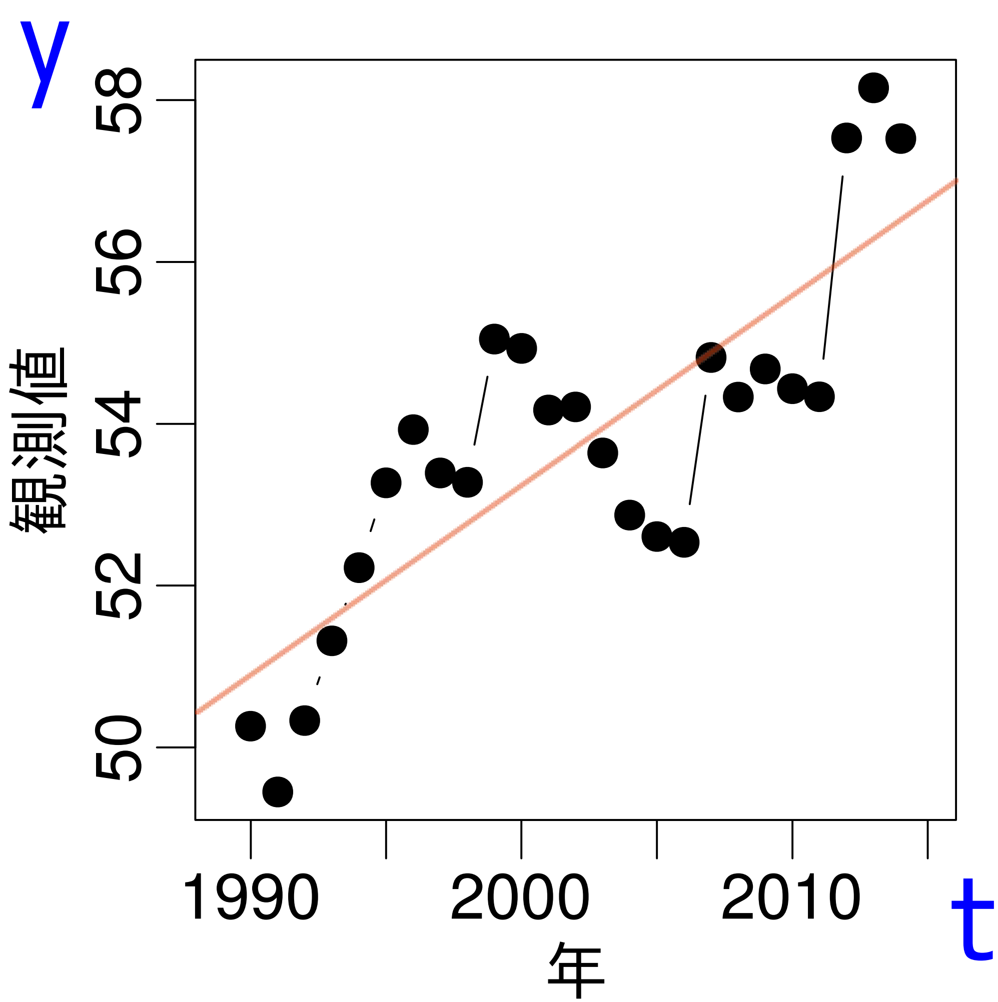
このような時系列データがあったとしましょう



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)

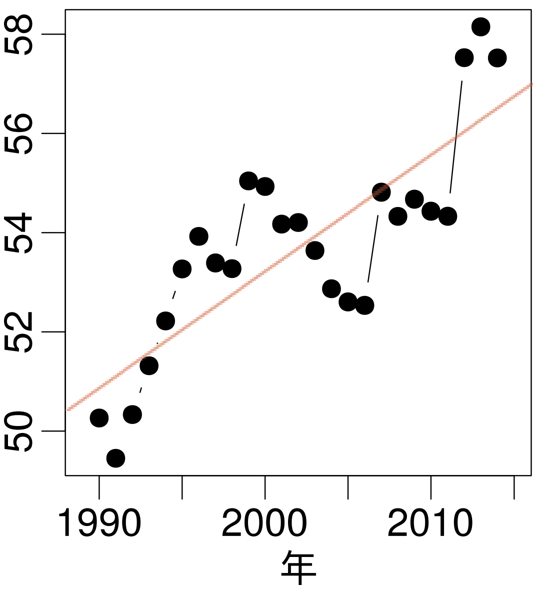
時系列データの統計モデリング入門



glm(y ~ t)

…とモデル
をあてはめてみた

「やったーゆーいだ!!」……??

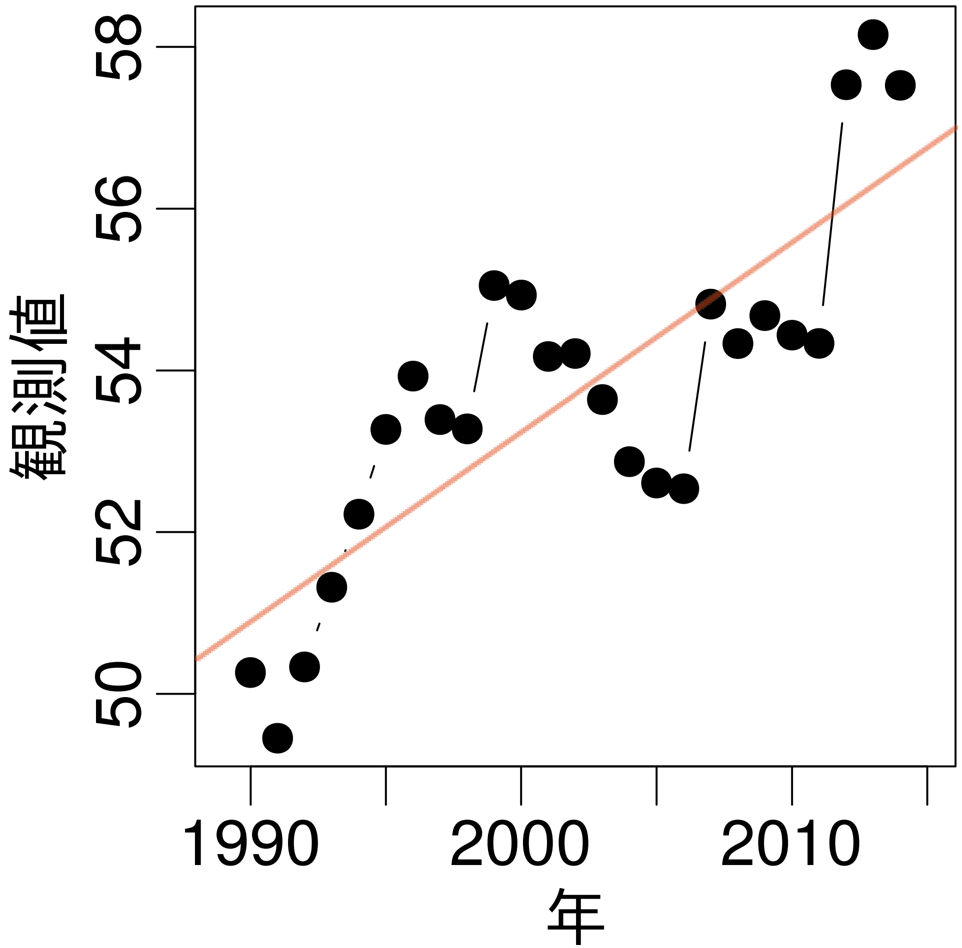



```
> summary(glm(formula = y ~ t))

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

time autocorrelation among data points!



「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

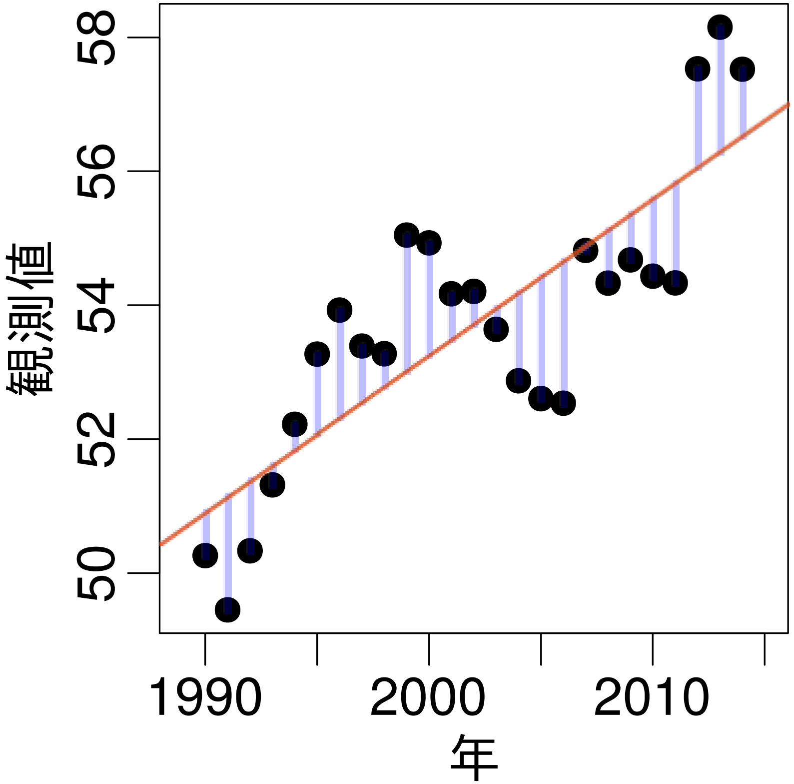
しても同様になる

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?

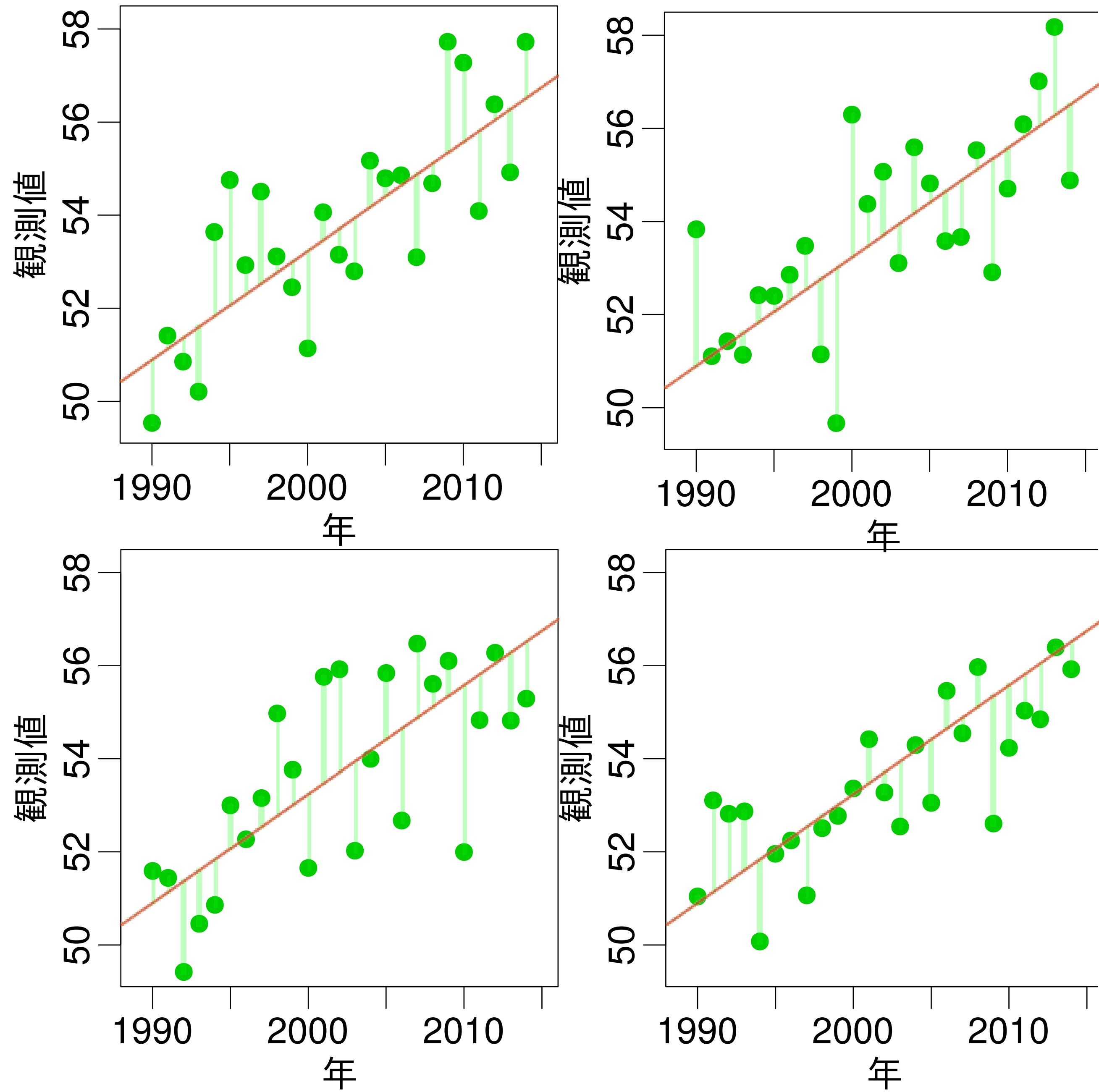
時系列の「ずれ」

auto-correlation



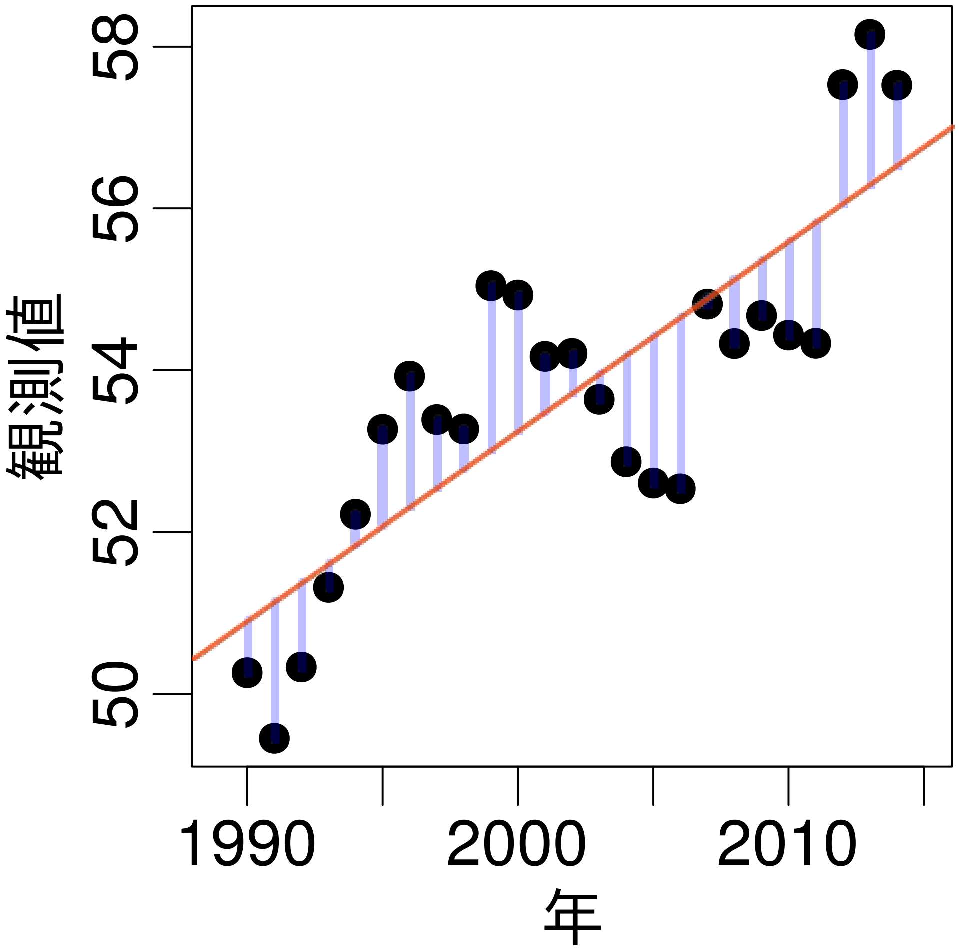
GLM のずれ

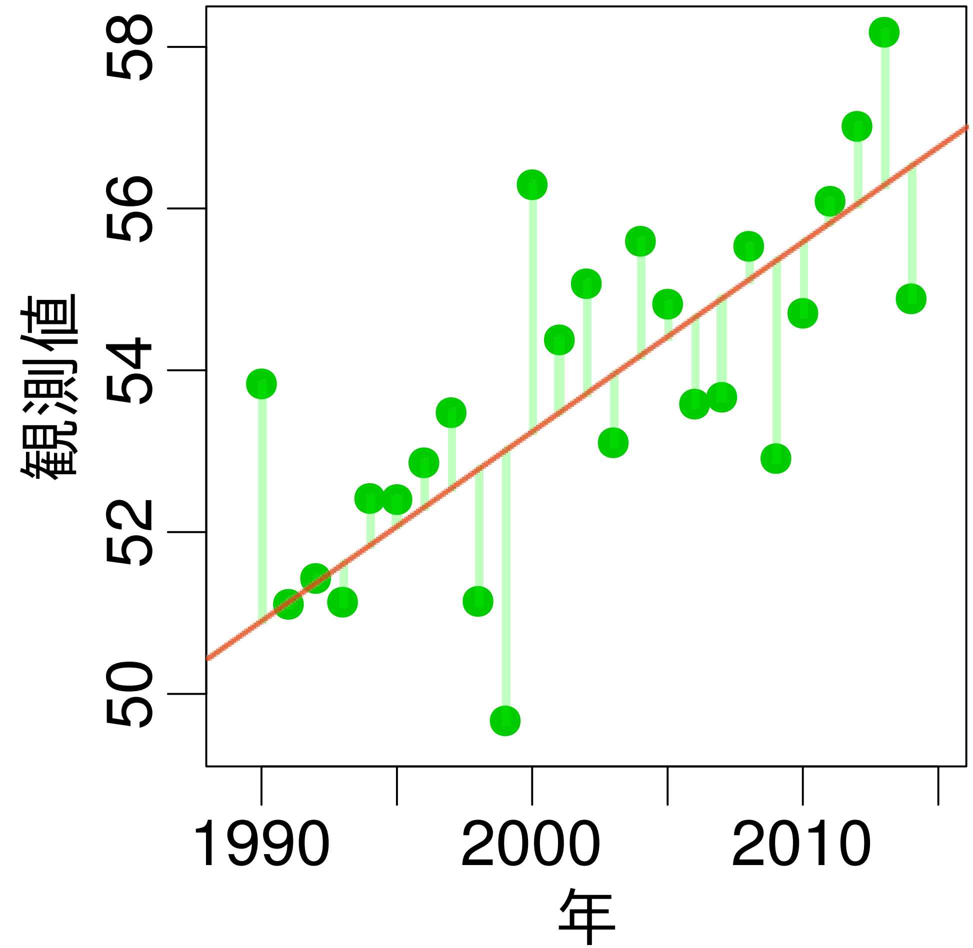
no correlation to adjacent points!



ずれかたが
ちがってる?

時系列の「ずれ」　　GLM のずれ



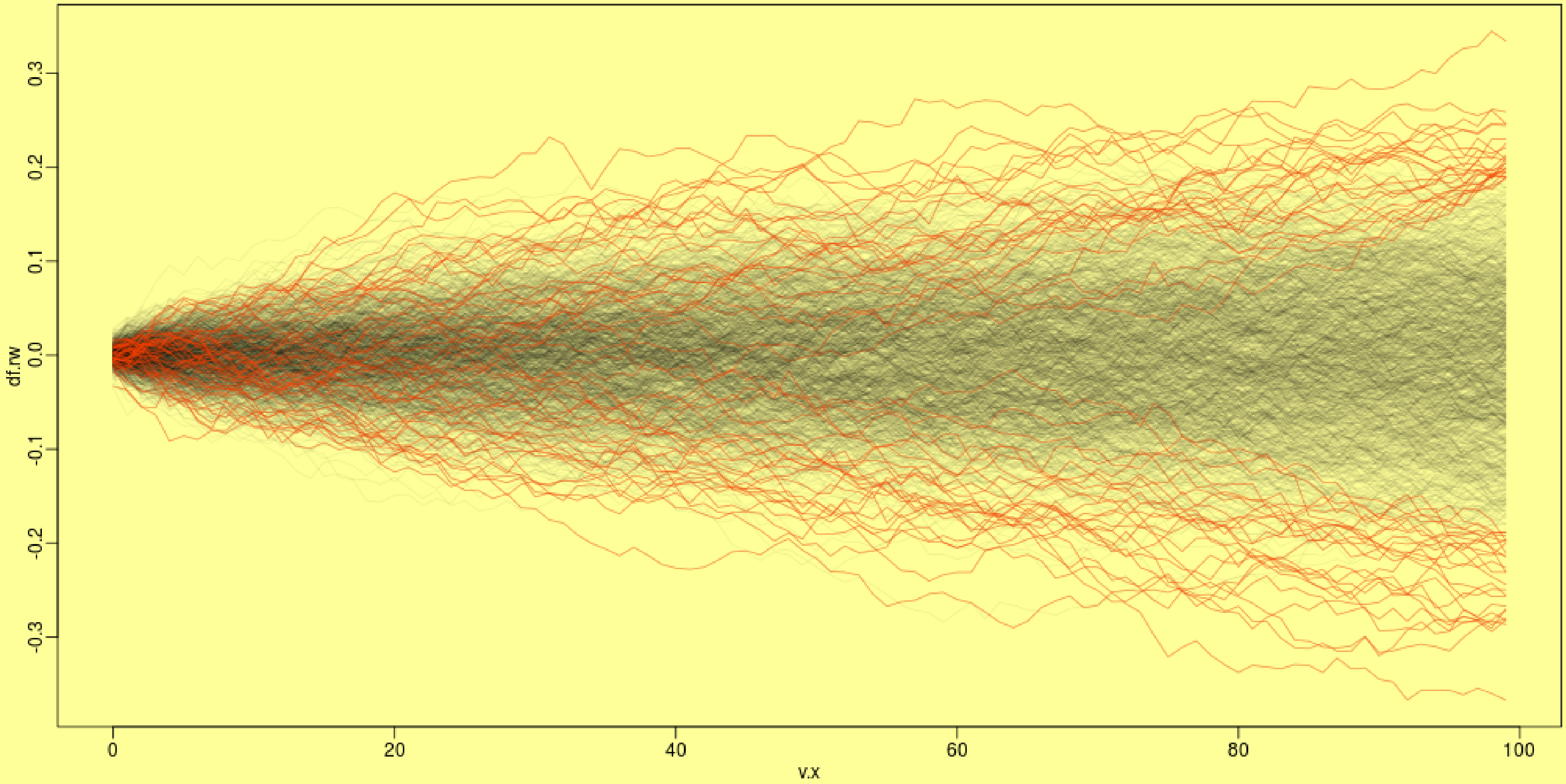


直線からのずれがちがう!

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

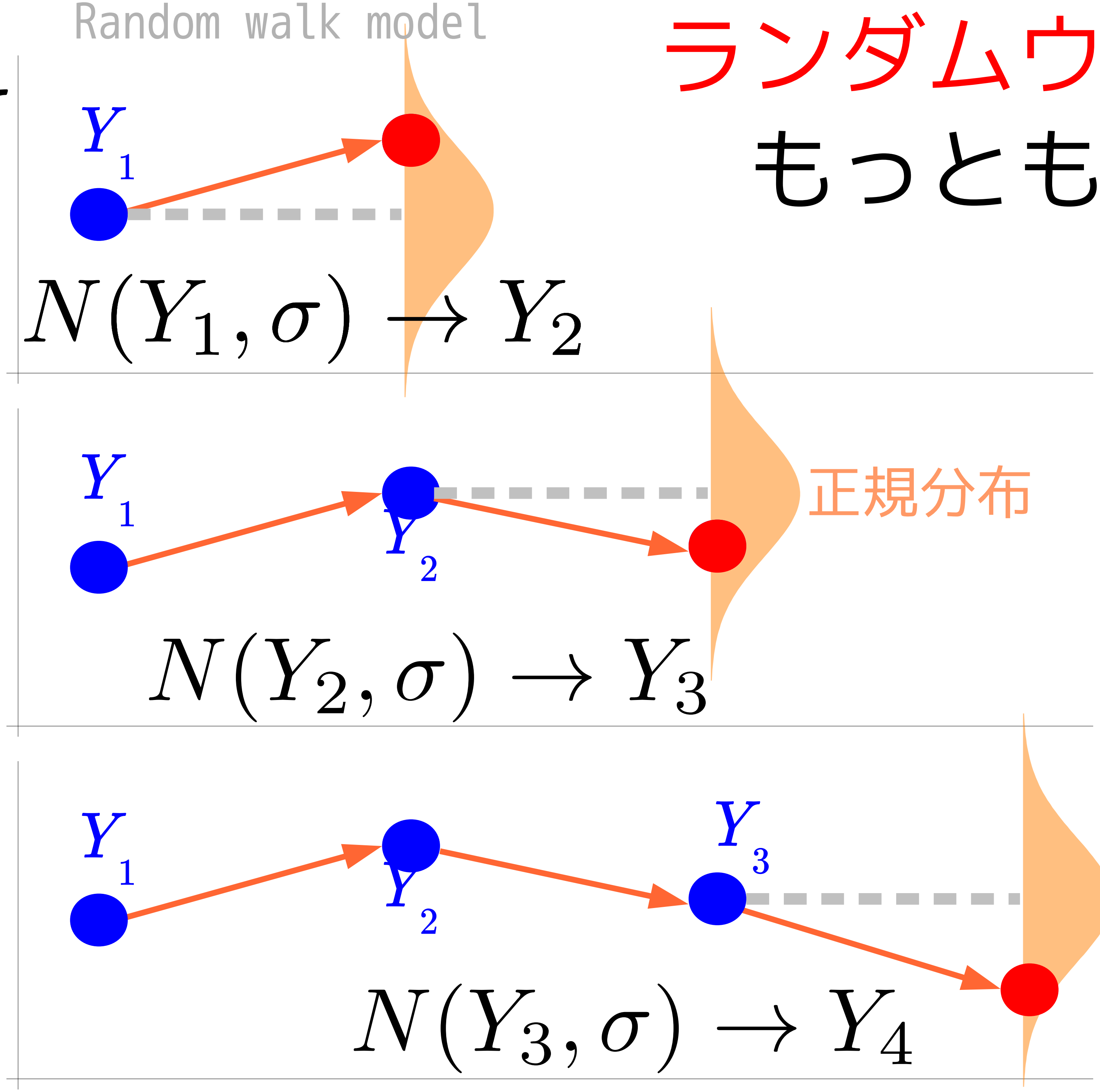
ランダムウォーク（乱歩）



Random walk model

ランダムウォーク

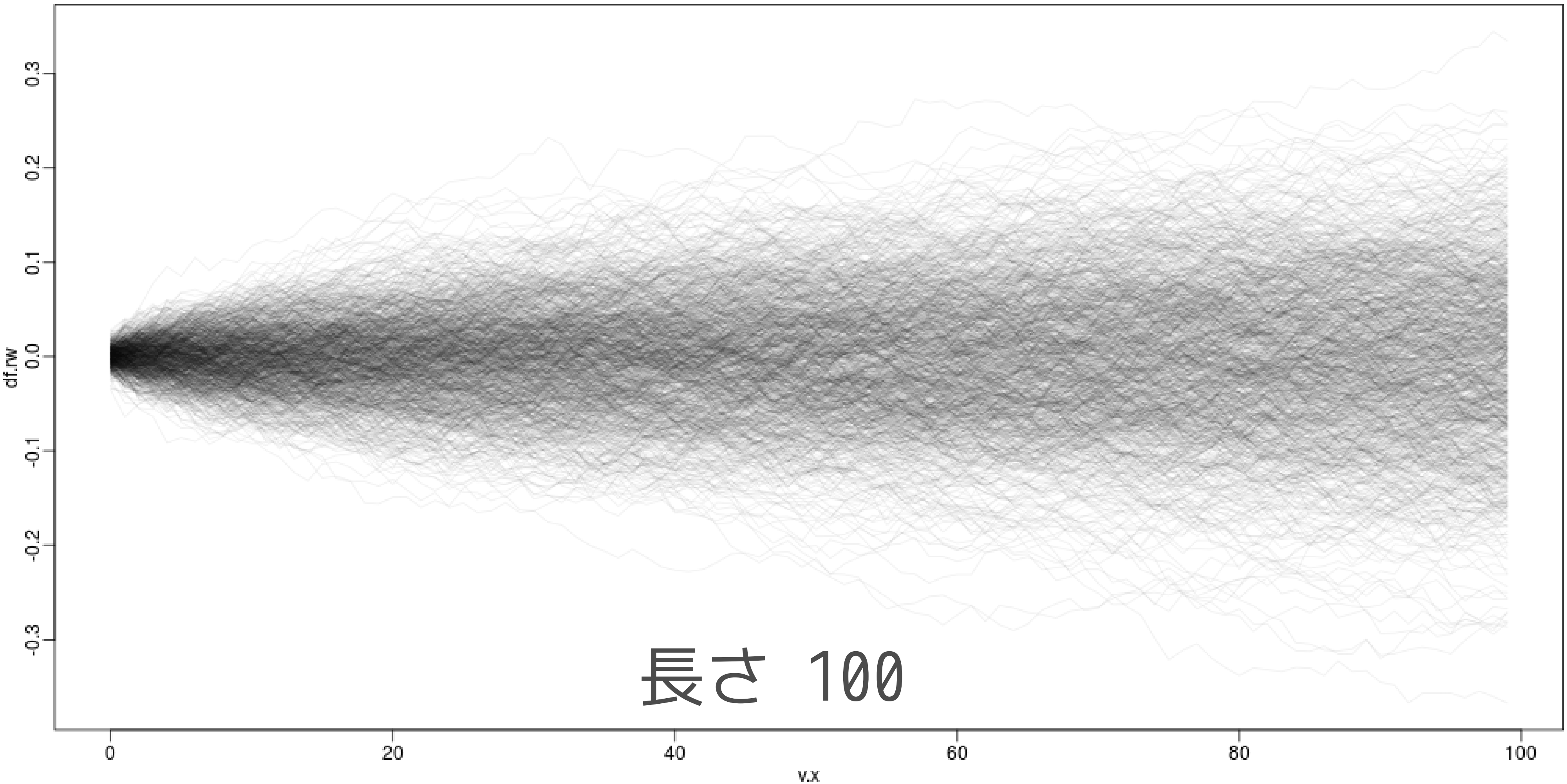
もっとも単純なモデル



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

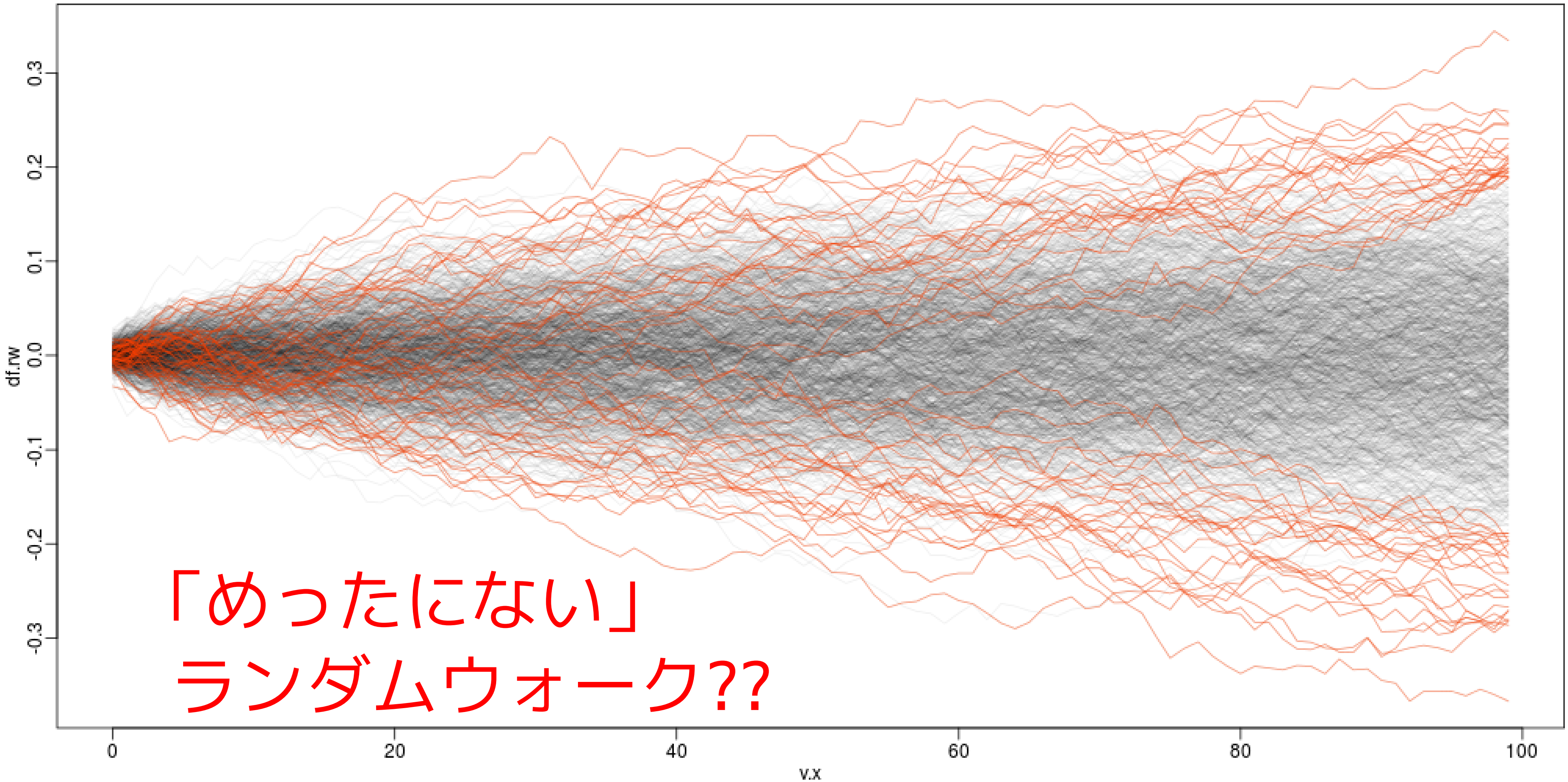
Generate 1000 time-series using random walk model



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

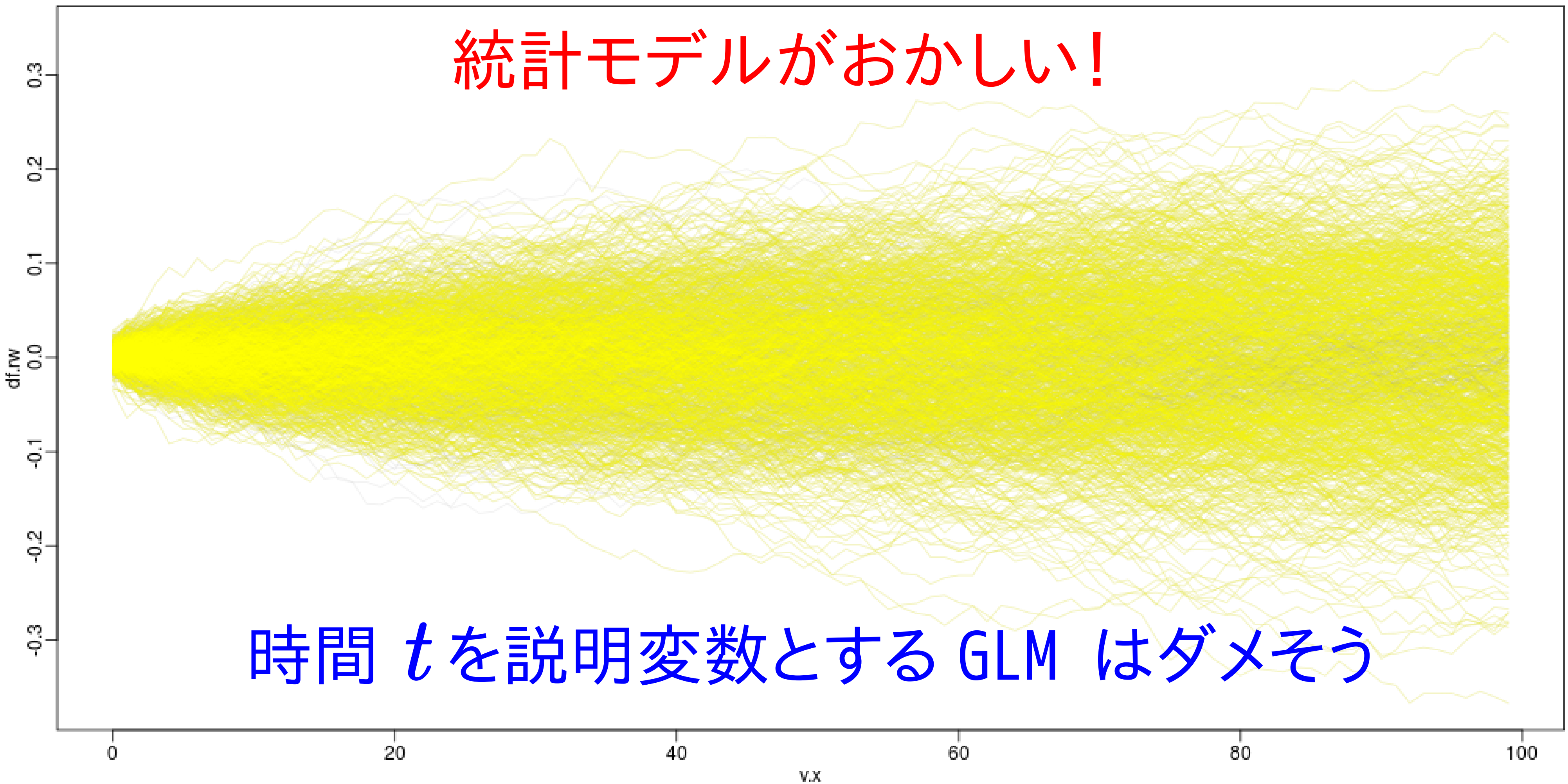
exceptional 50 time-series data?



しかし直線回帰 GLM あてはめると…

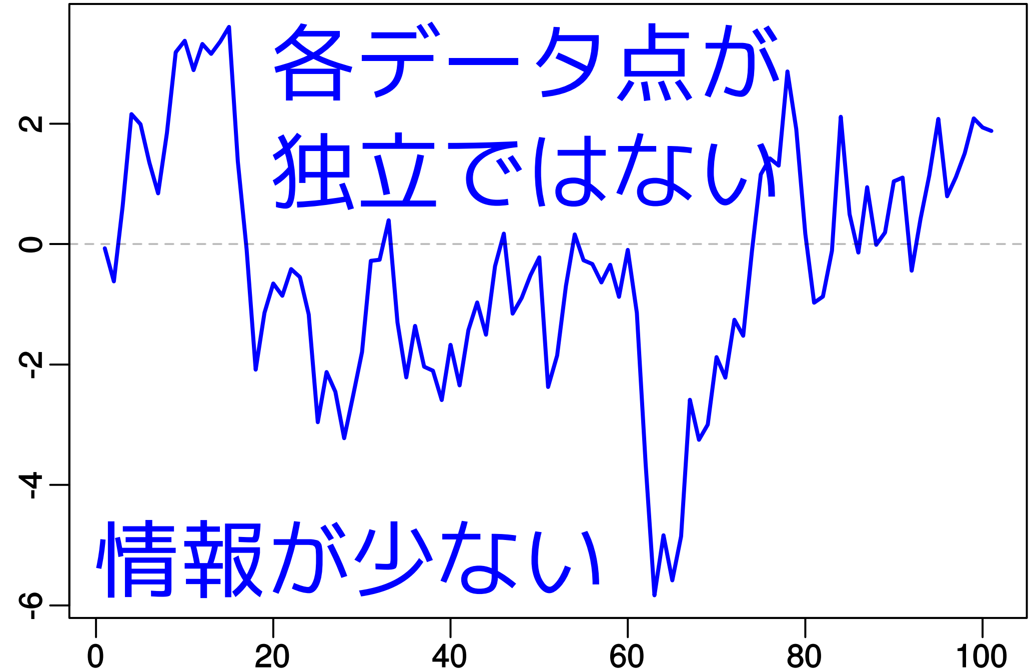
ほとんどすべての場合で「ゆーい」!

significant? no!

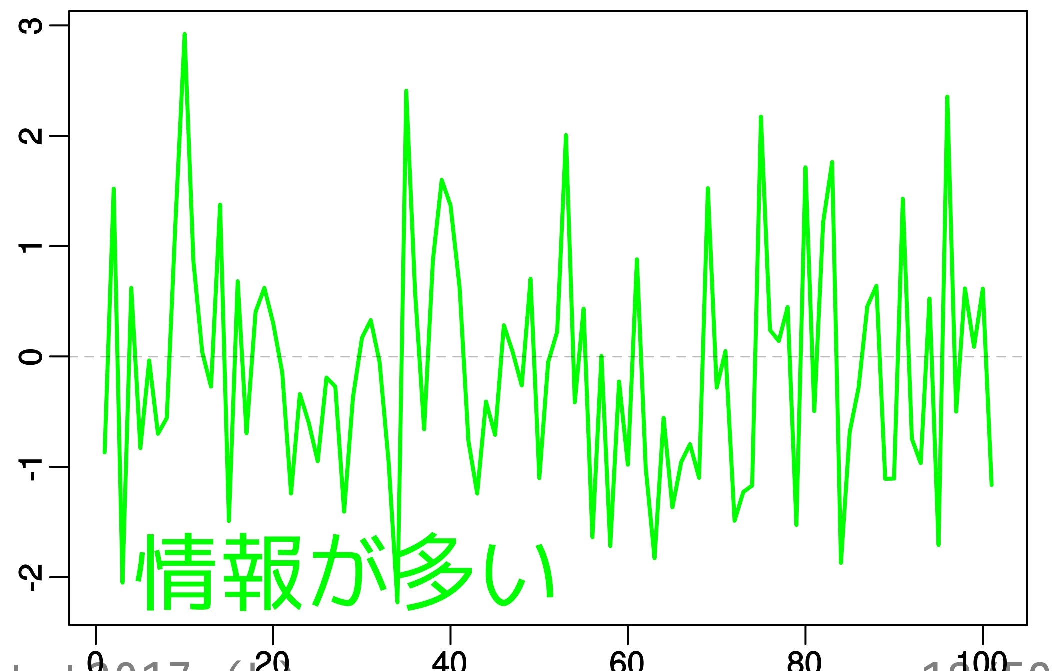


ちょっとでも傾いてたら「ゆーい」

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



temporal auto-correlation coefficient

時間的自己相関

(略称:自己相関，時間相関)

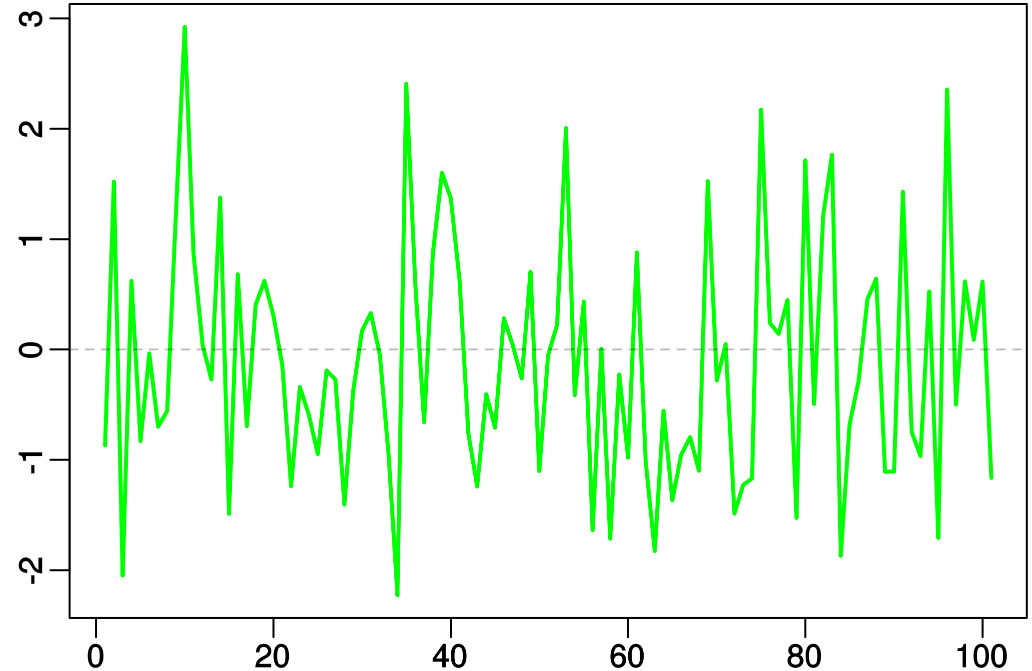
を調べたらいいの?

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

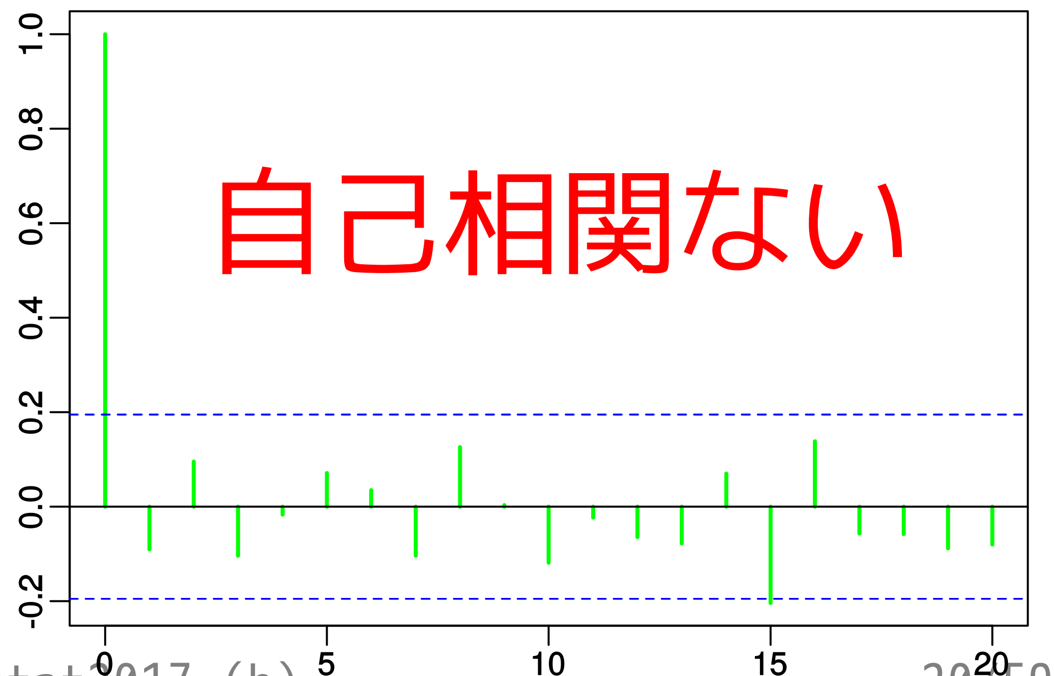
R の ts クラス: 時系列をあつかう

plot(ts(Y))

これはたんなる
100 個の正規乱数

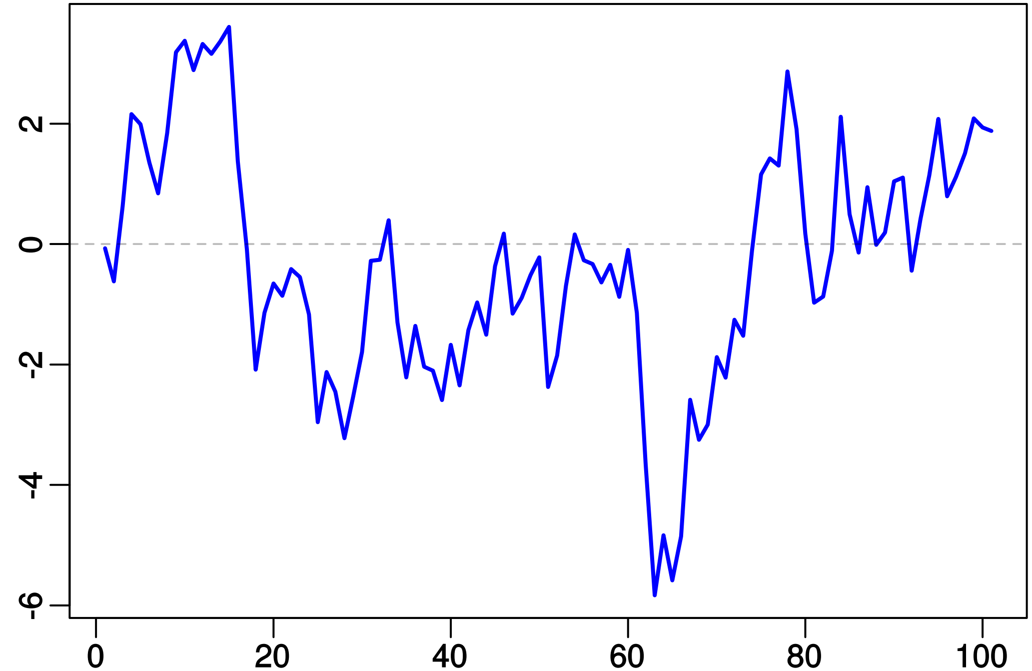


plot(acf(ts(Y)))

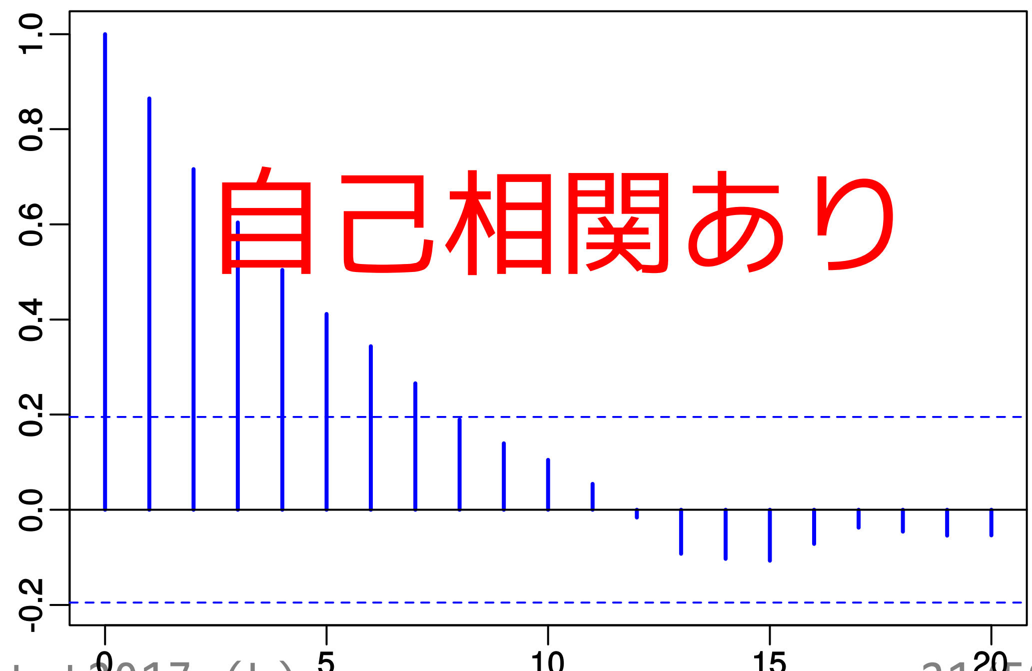


自己相関減衰の様子を図示

plot(ts(Y))

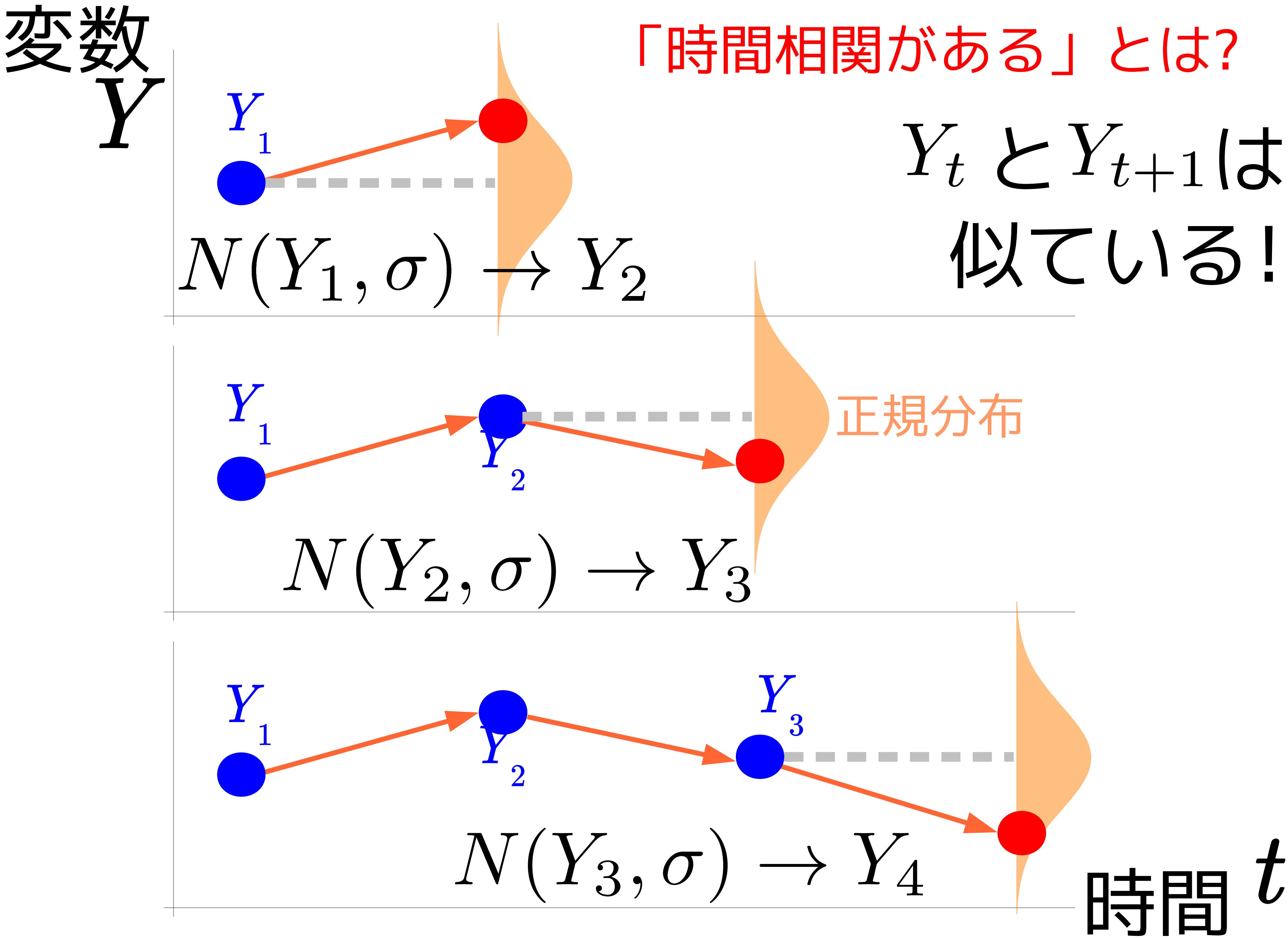


plot(acf(ts(Y)))



「時間相関がある」とは?

Y_t と Y_{t+1} は似ている!



temporal auto-correlation coefficient

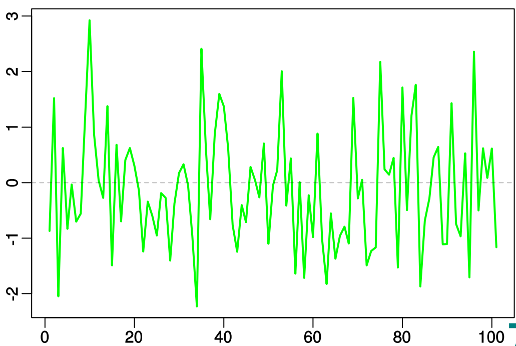
時間的自己相関

いつも役にたつわけではない?

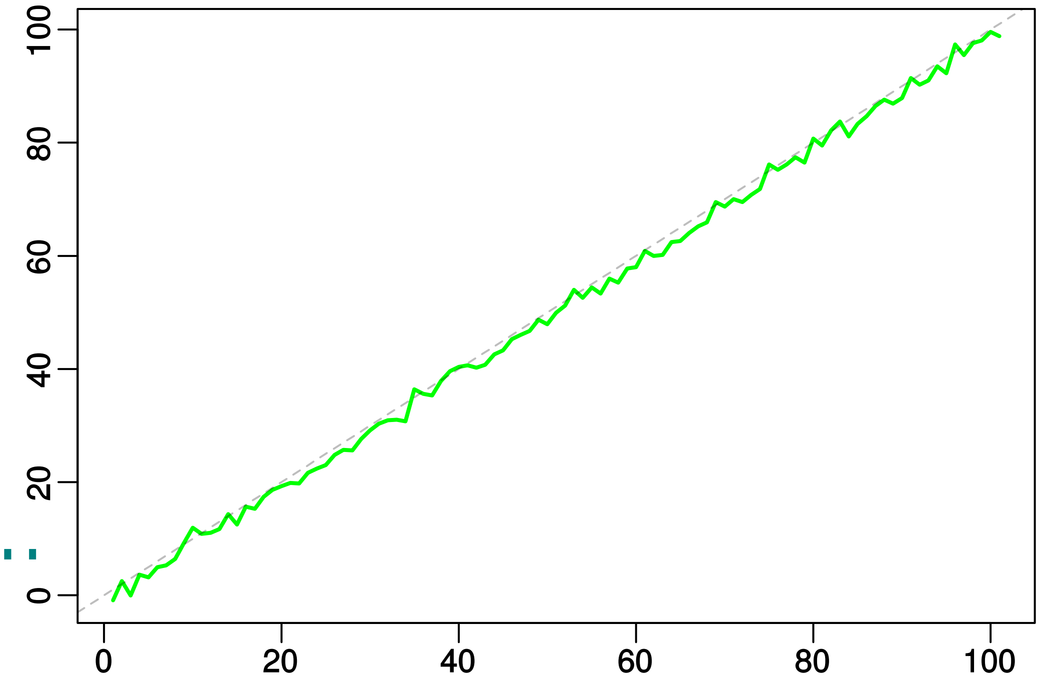
$$\rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-1})}}$$

各点独立のデータをナナメにすると?

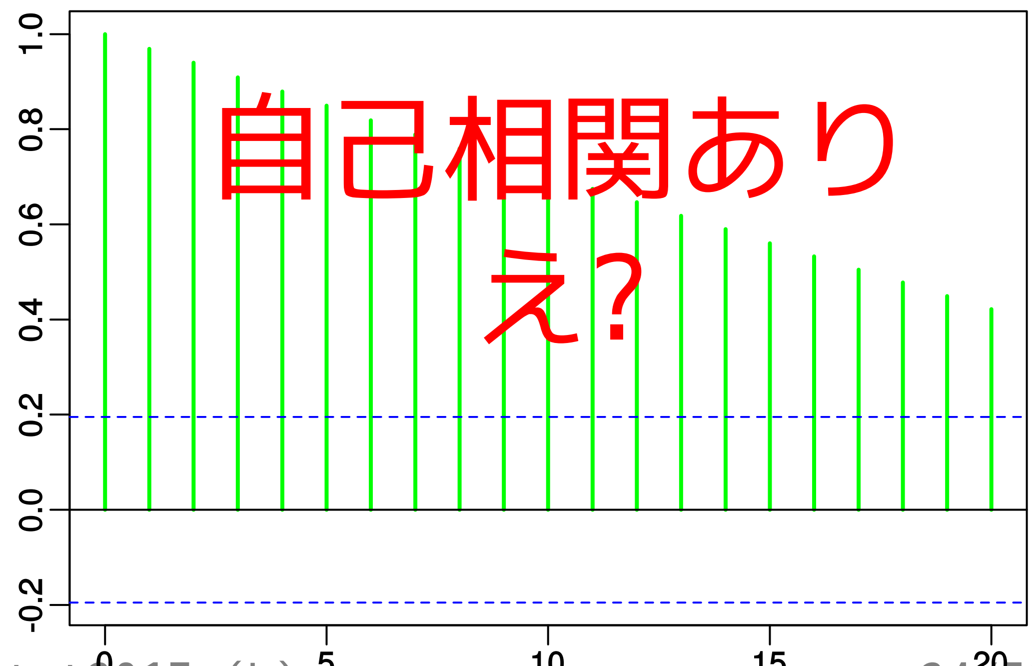
plot(ts(Y))



これを
ナナメに
したもの
なんだけど…

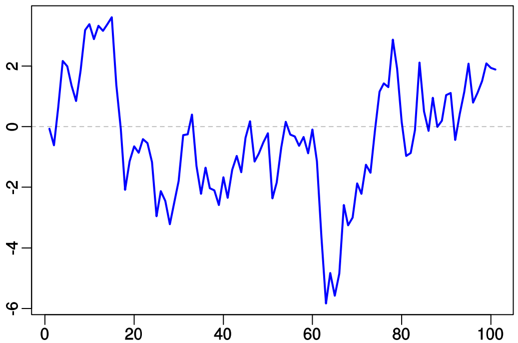


plot(acf(ts(Y)))

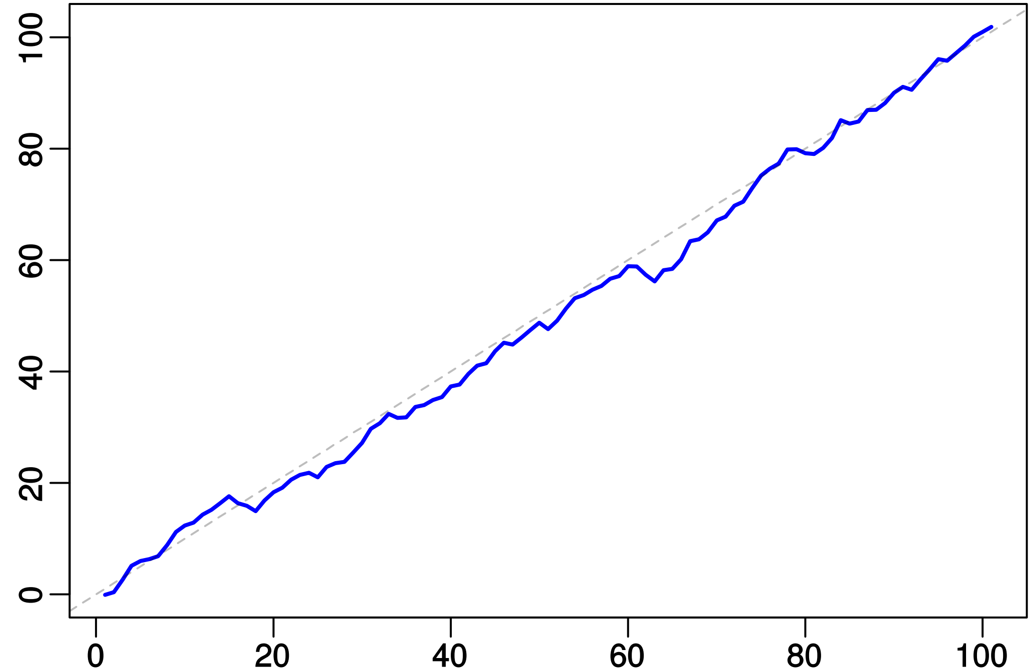


各点独立のデータをナナメにすると?

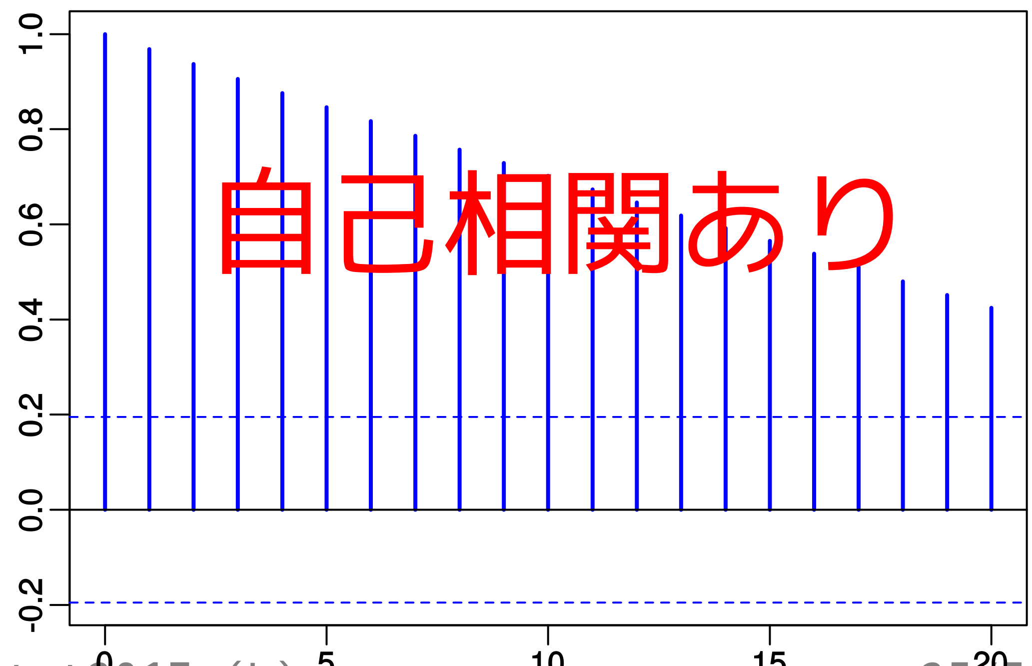
plot(ts(Y))



これを
ナナメに
したもの

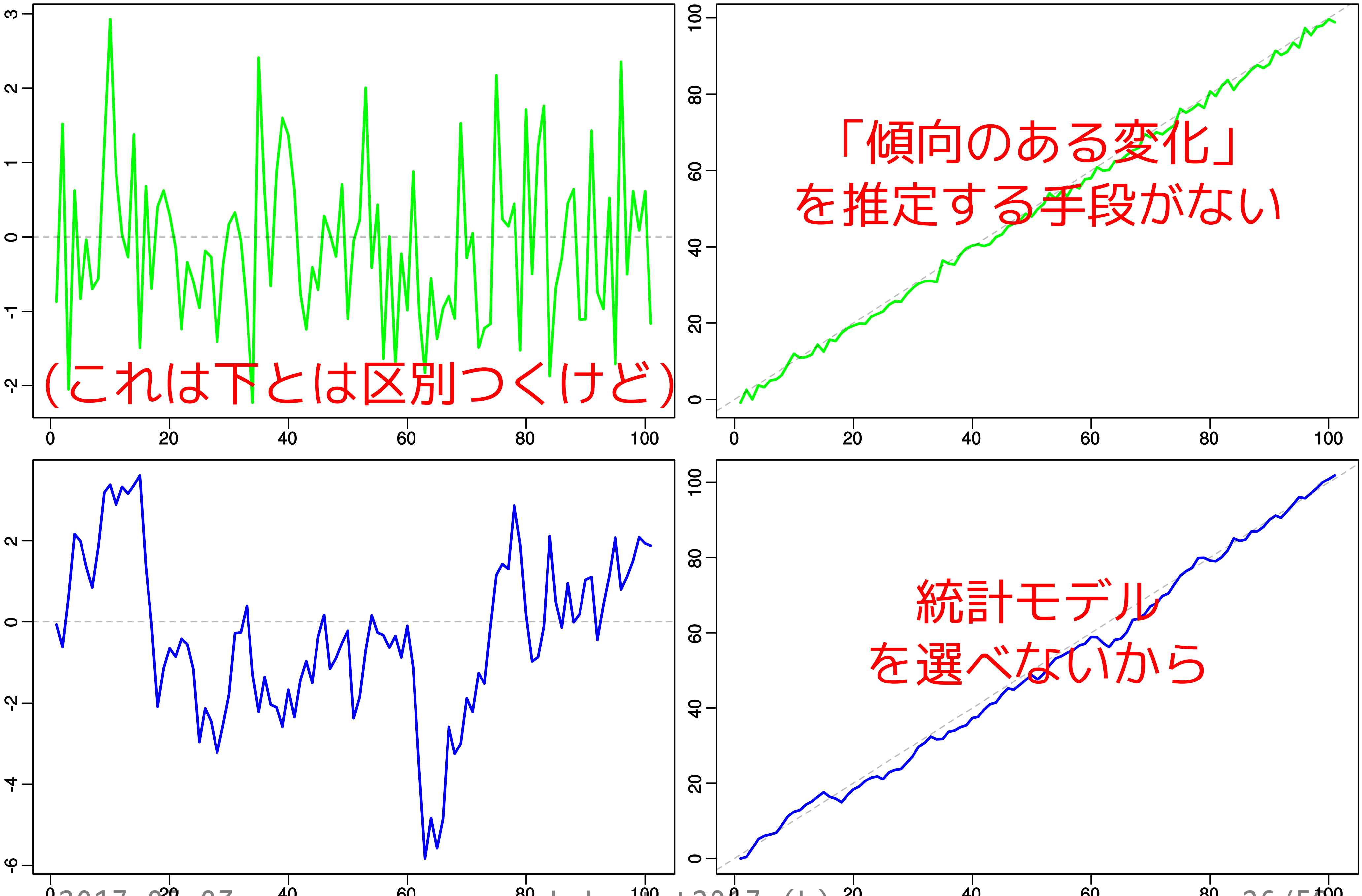


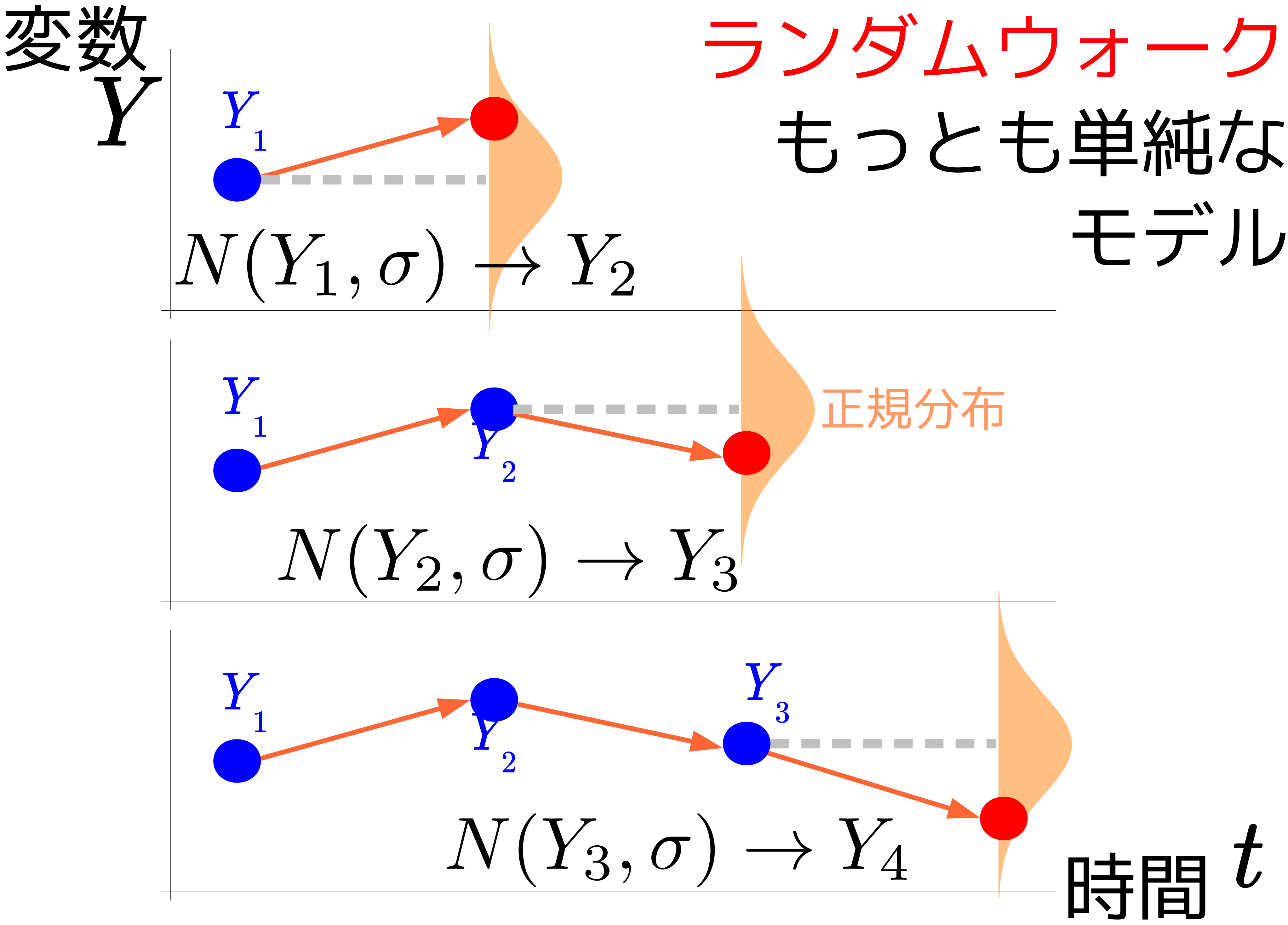
plot(acf(ts(Y)))



自己相関あり

自己相関係数みても区別がつかない





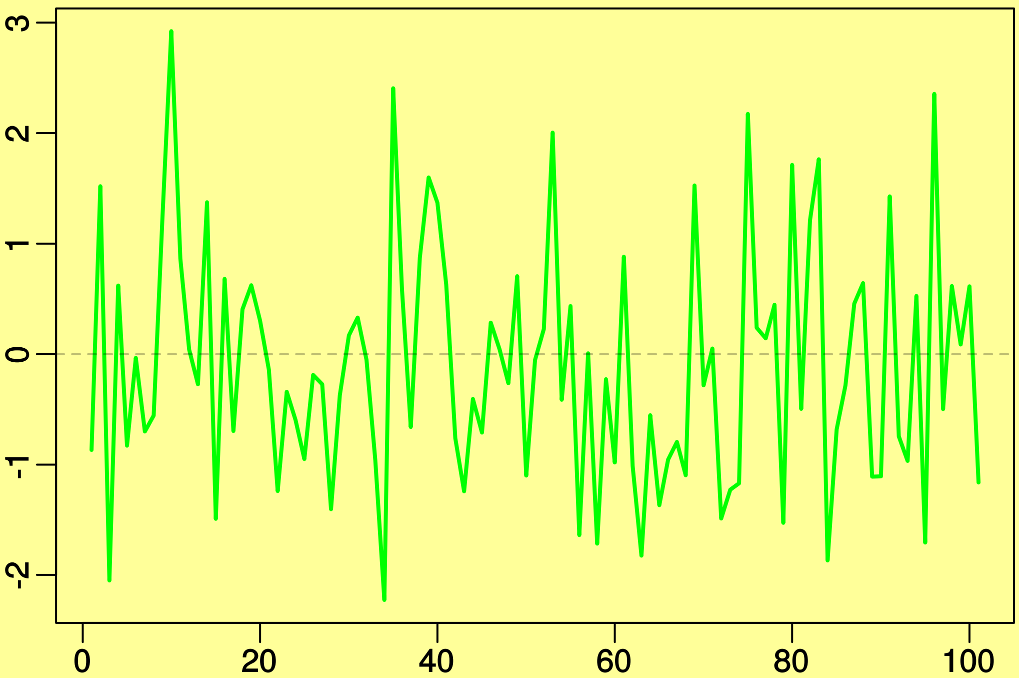
変数
Y
ランダムウォーク
もっとも単純な
モデル
Y_1
$N(Y_1, \sigma) \to Y_2$
Y_1
Y_2
正規分布
$N(Y_2, \sigma) \to Y_3$
Y_1
Y_2
Y_3
$N(Y_3, \sigma) \to Y_4$
時間 t

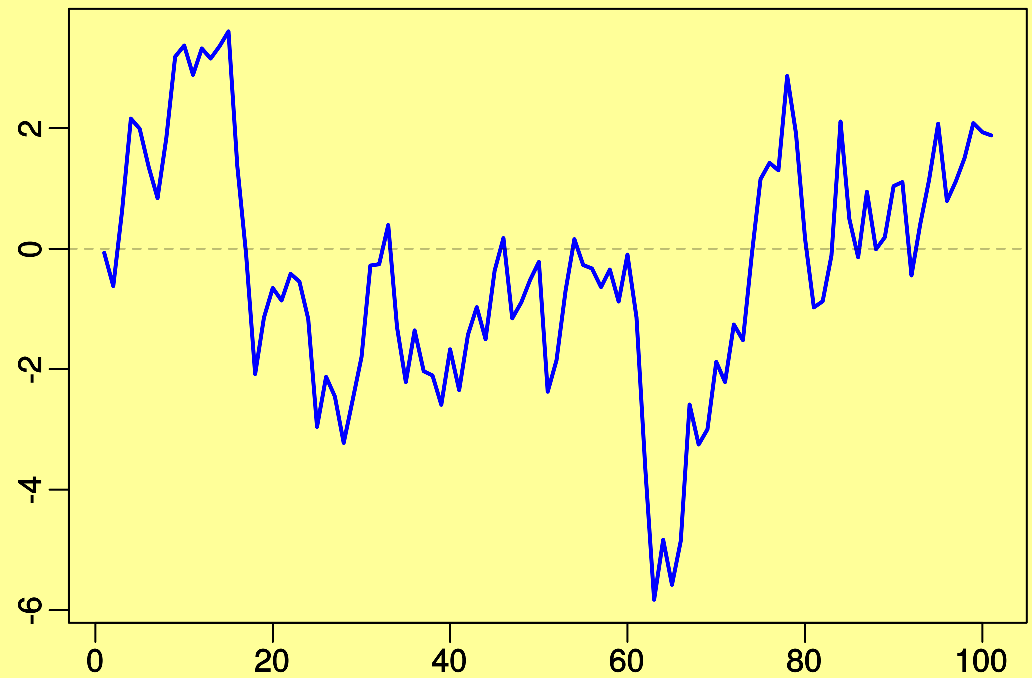
状態空間モデルでたちむかう

時系列データ解析

いろいろな時系列データを

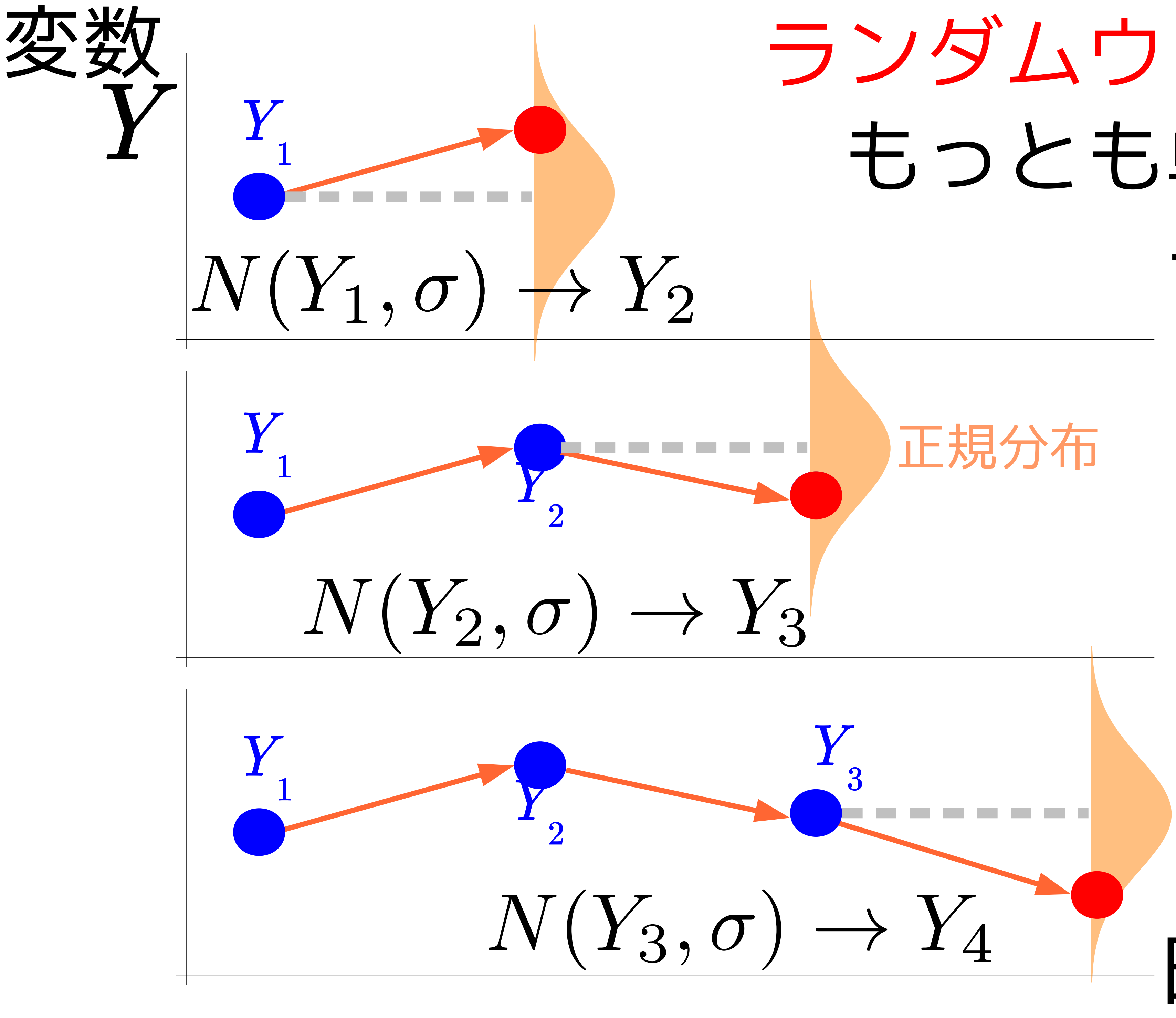
統一的にあつかえないか?





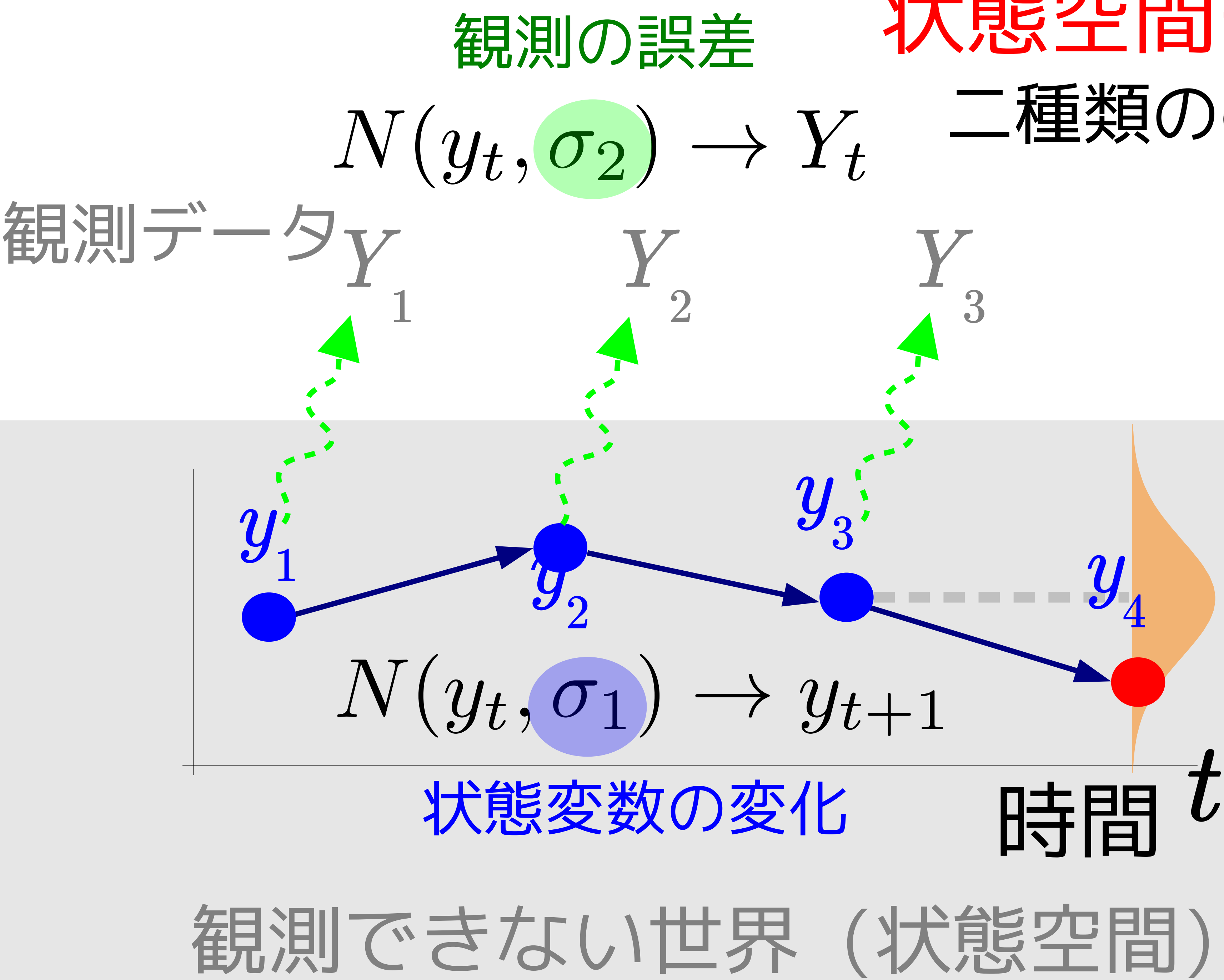
ランダムウォーク

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

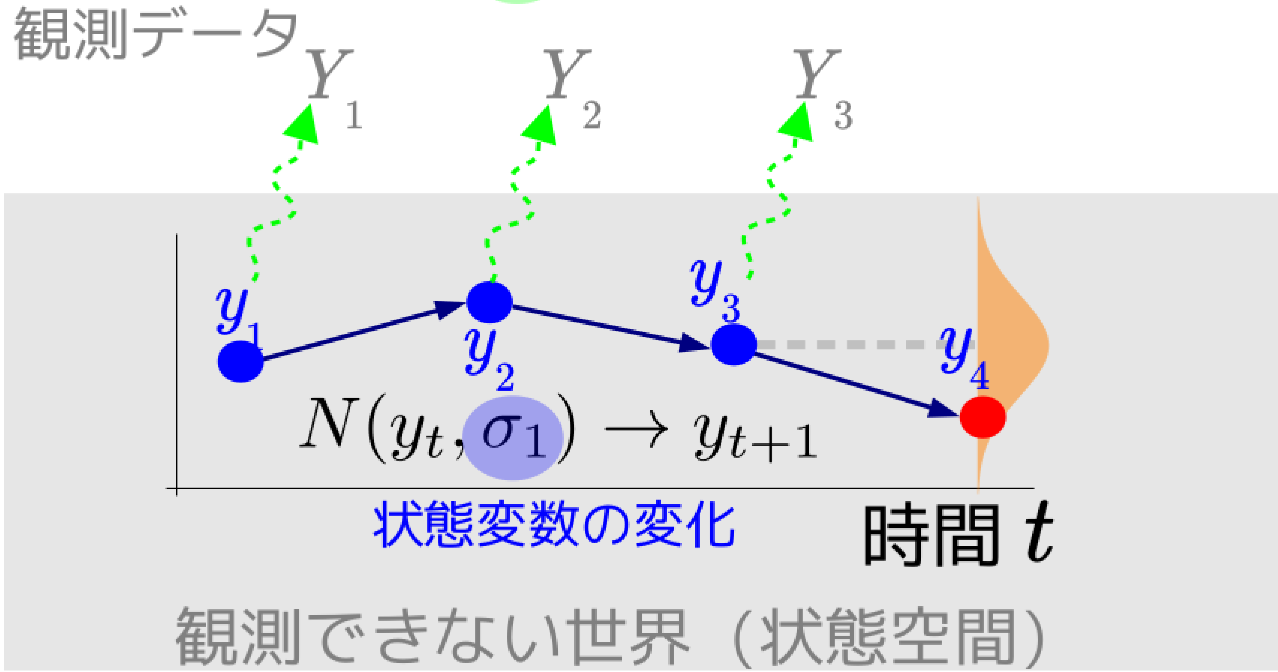


State-space model!

観測の誤差

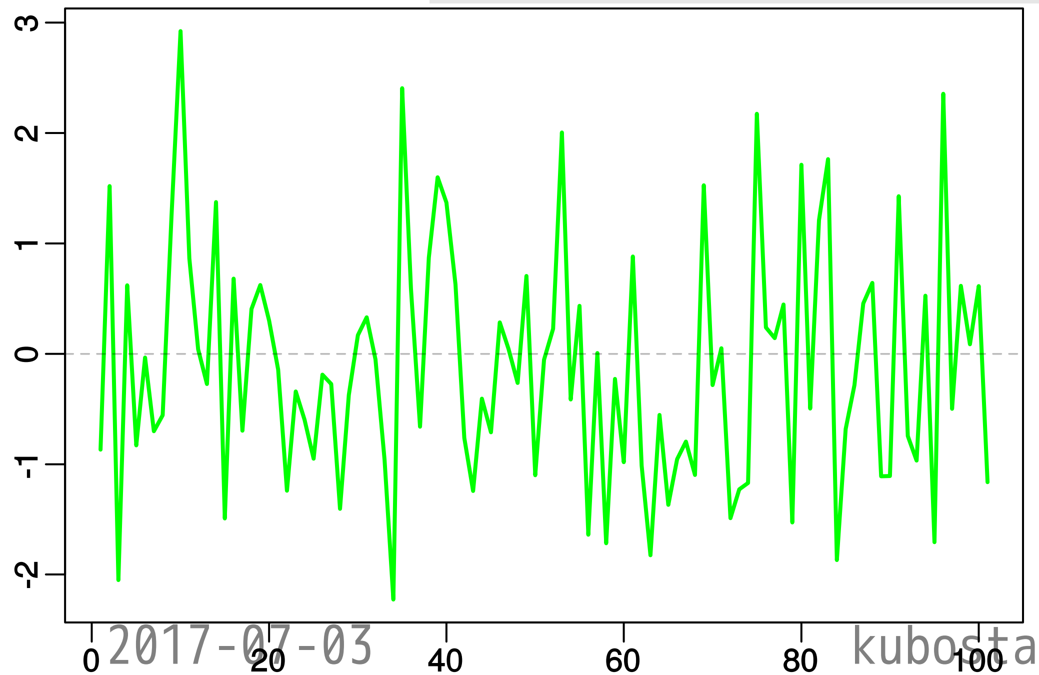
状態空間モデル

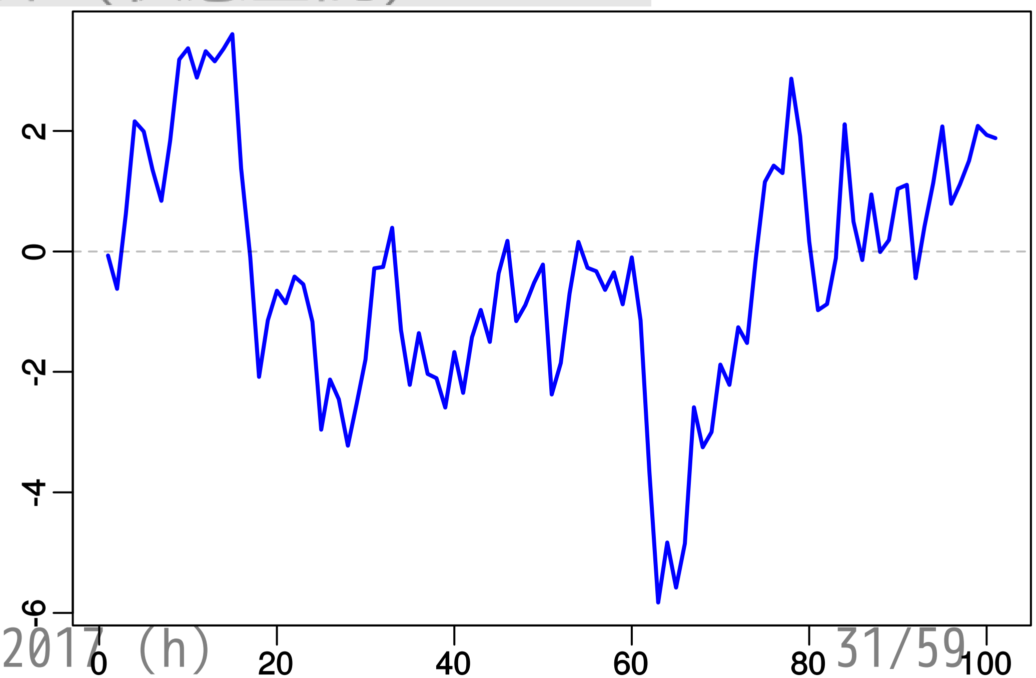
$N(y_t, \sigma_2) \rightarrow Y_t$ 二種類のσをもつ



σ_2 大
σ_1 小

σ_2 小
σ_1 大





状態空間モデルは…

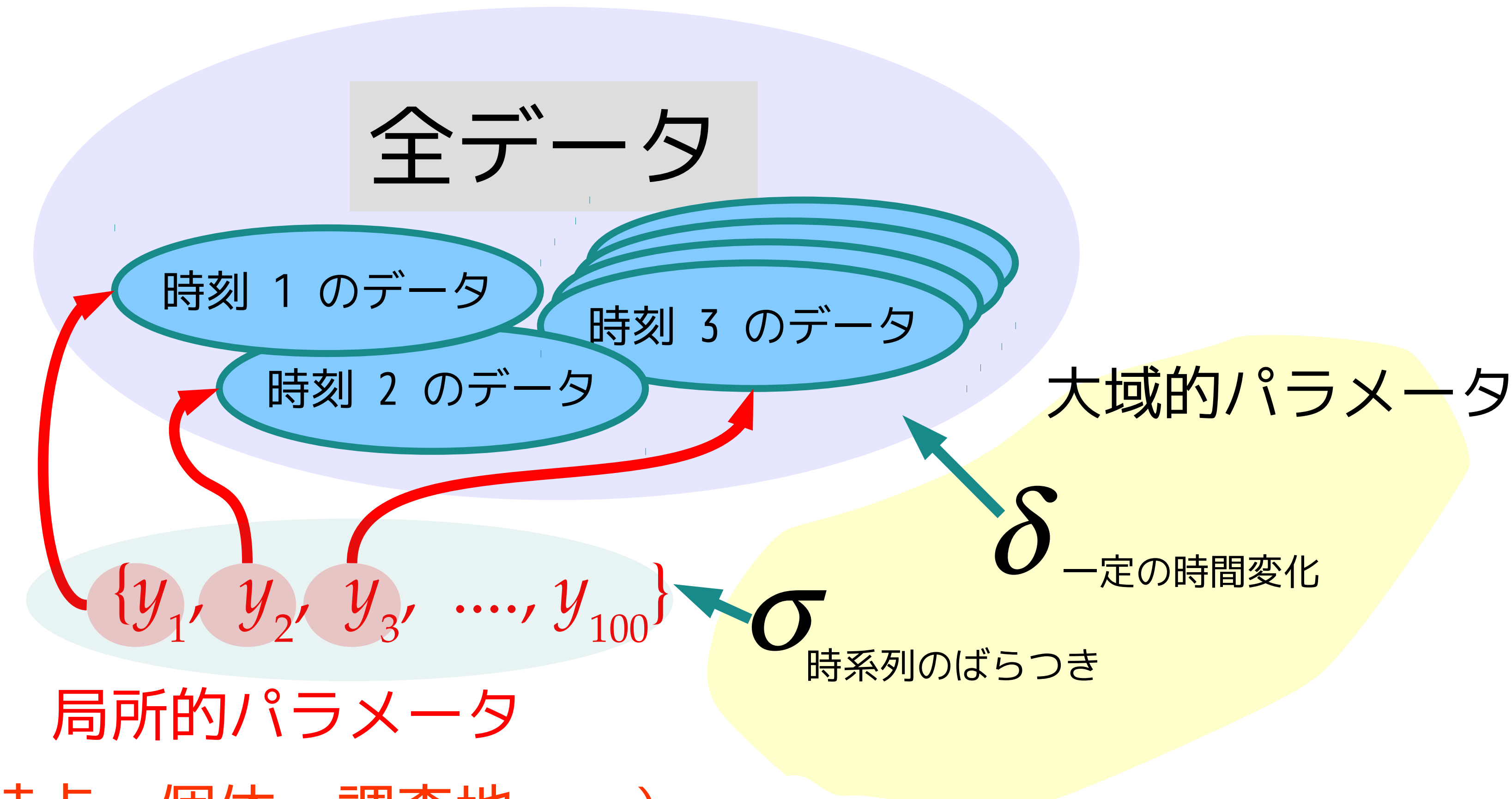
state-space model is ...

階層ベイズモデルだ!

a hierarchical Bayesian model!

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)
```

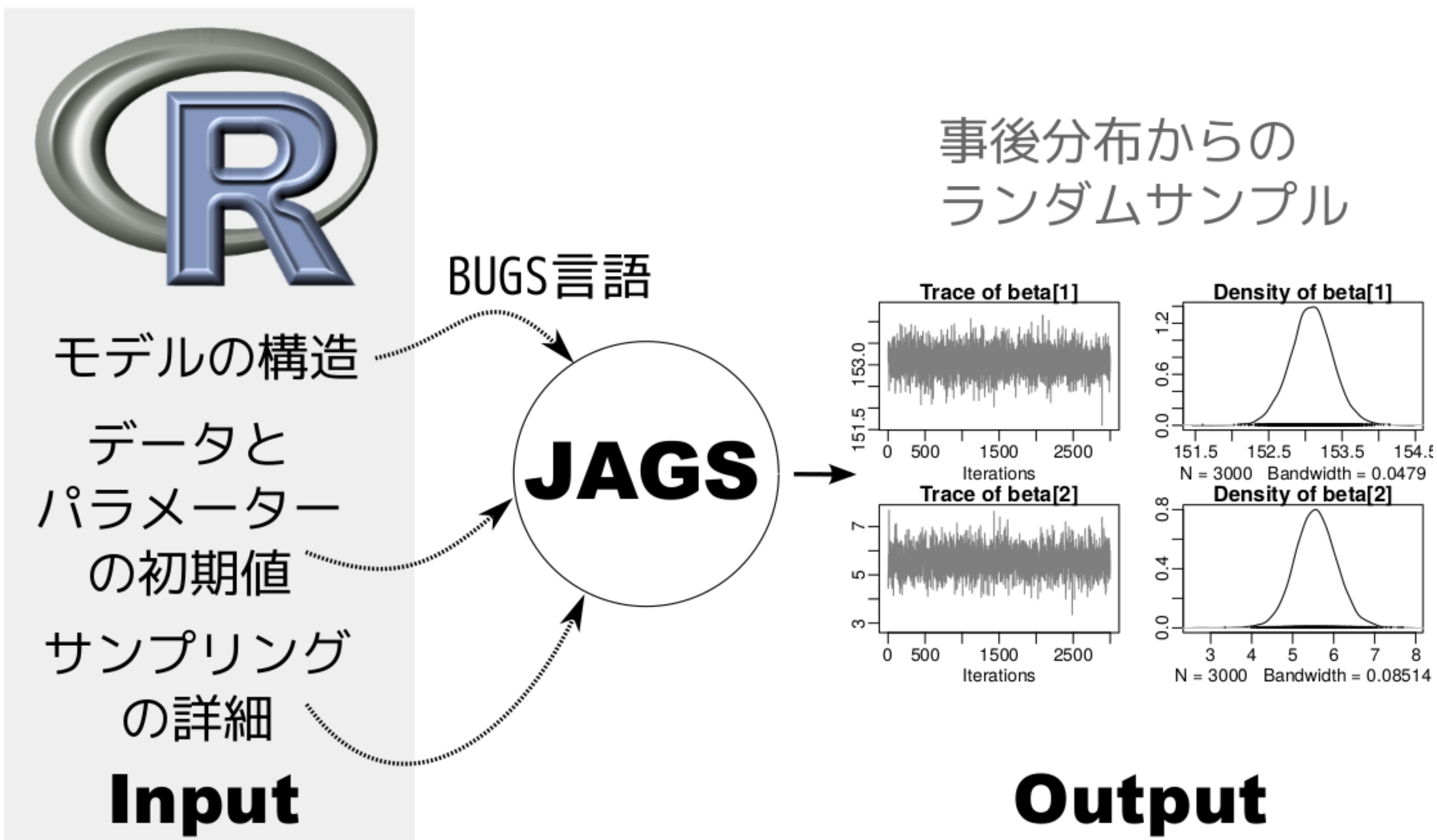
しかしより一般化したモデルに

ついての理解が必要かも

こういう問題も 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)
    }
}
```

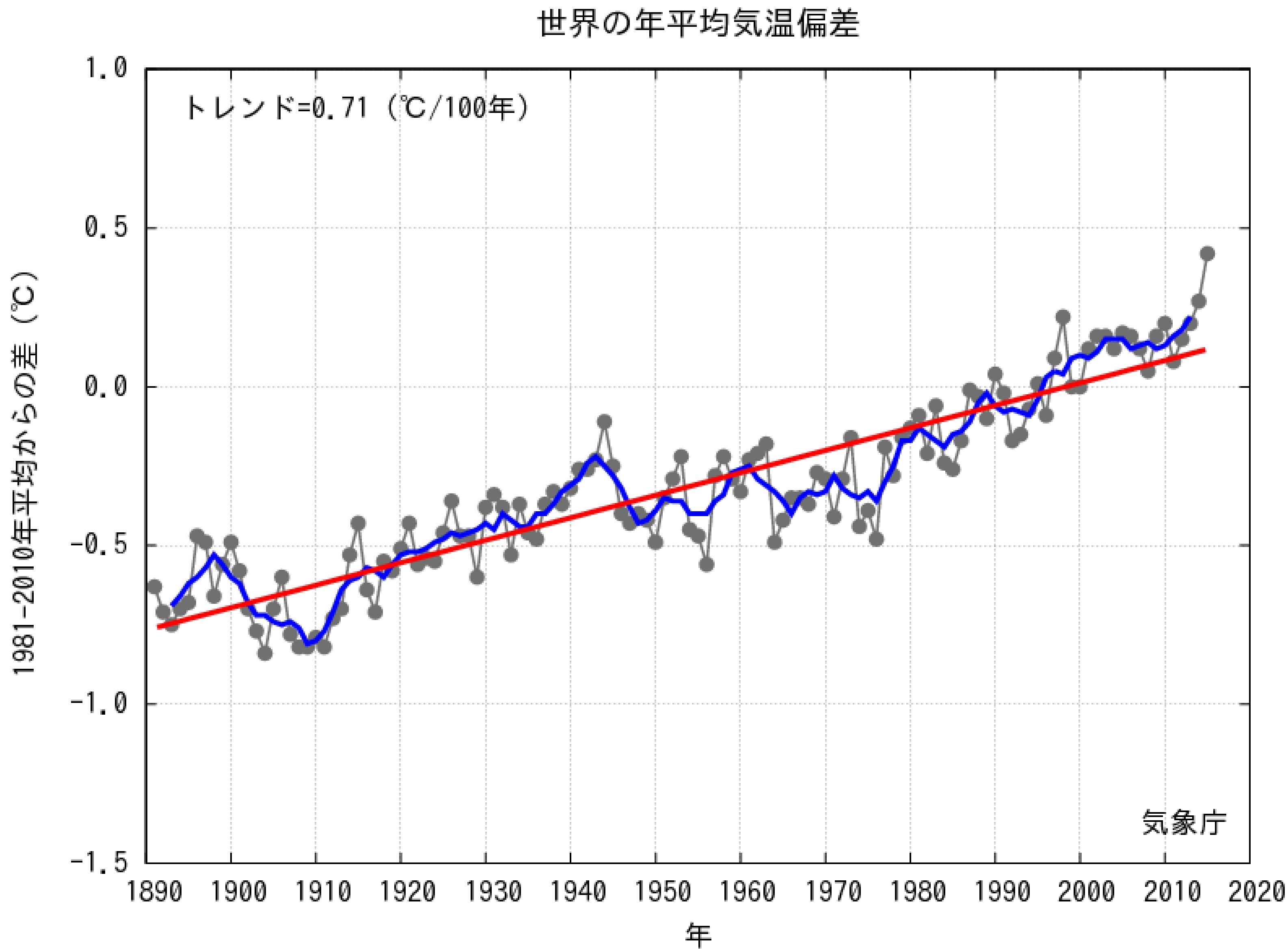
状態空間モデルを使う利点

「ばらばら解析」の回避

気象庁のデータ解析？

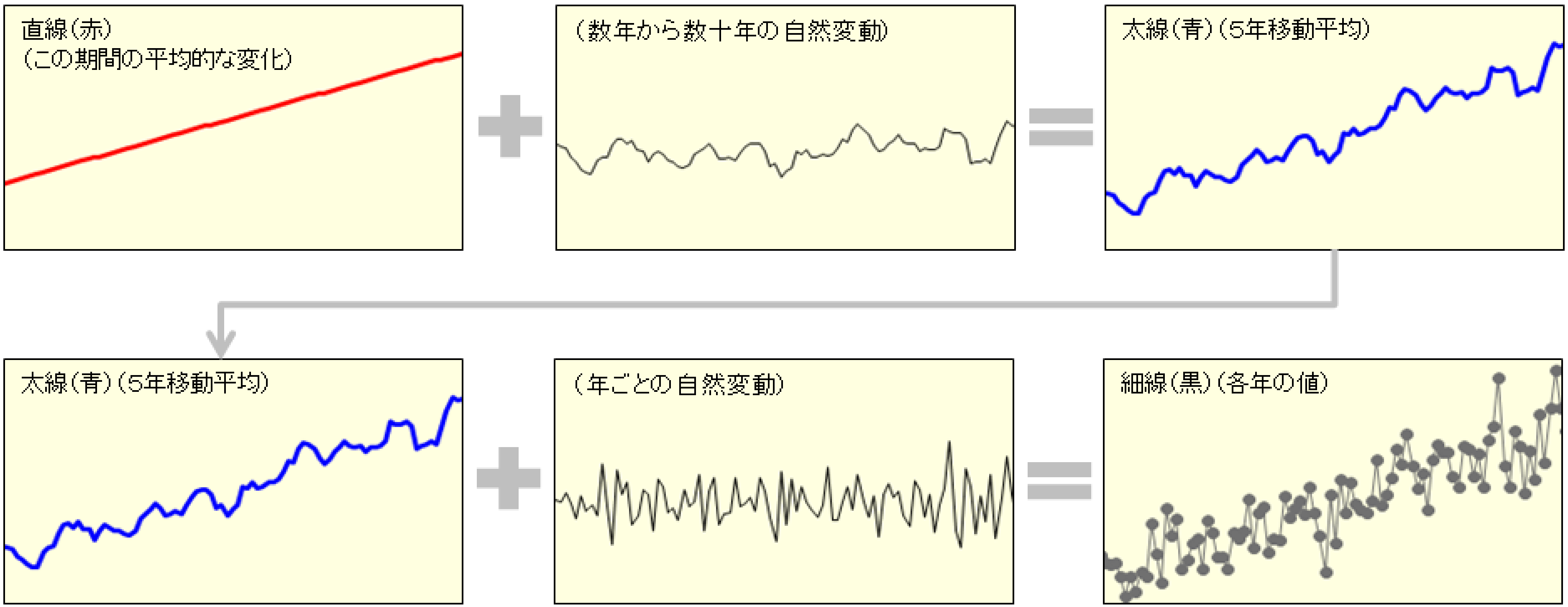
An example: time change of yearly temperature long-term change of yearly temperature

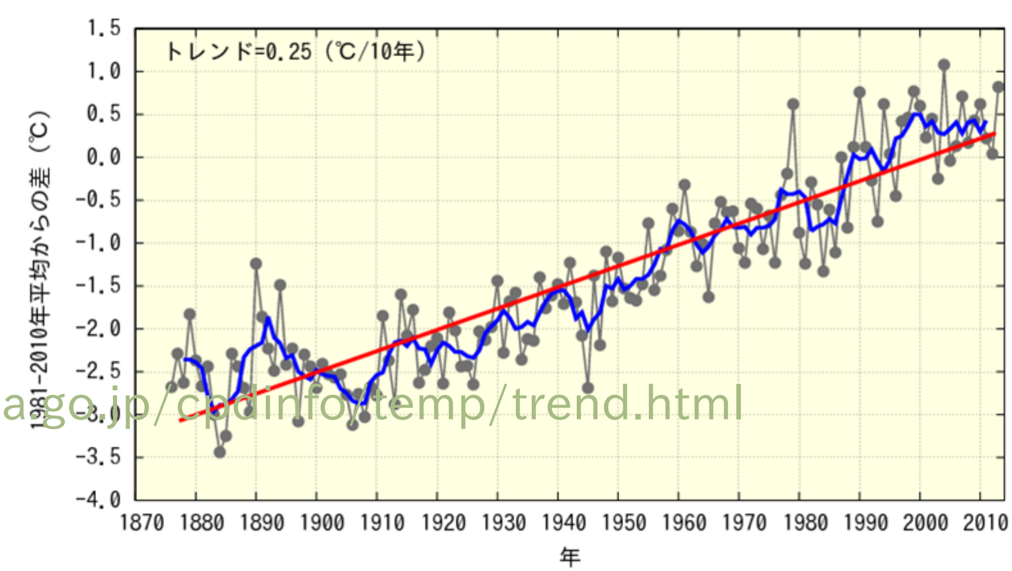
気象庁の長期変化傾向（トレンド）の解説



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

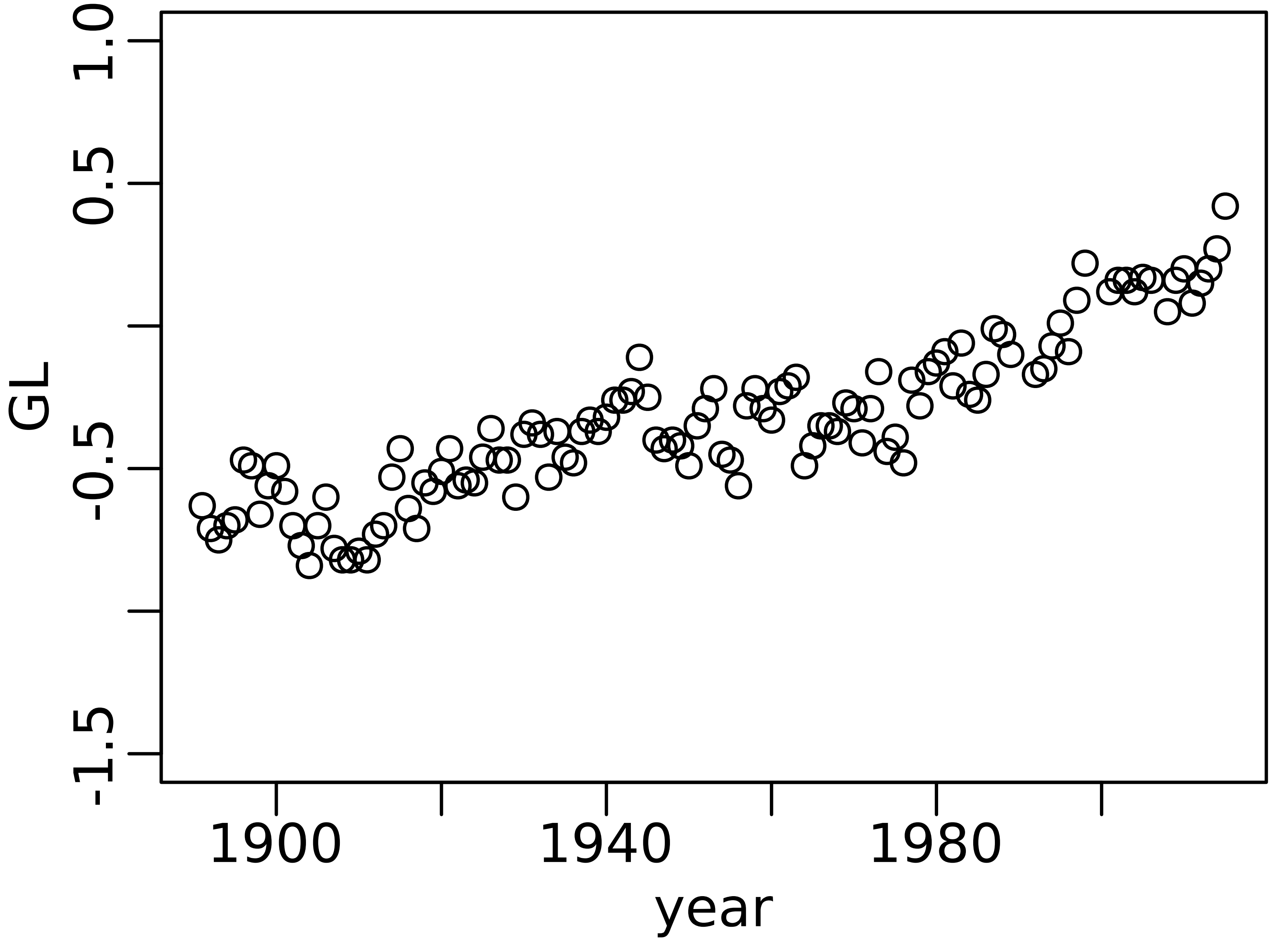




http://www.data.jma.go.jp/cpdinfo/temp/trend.html

2016-08-09

公開データをダウンロード

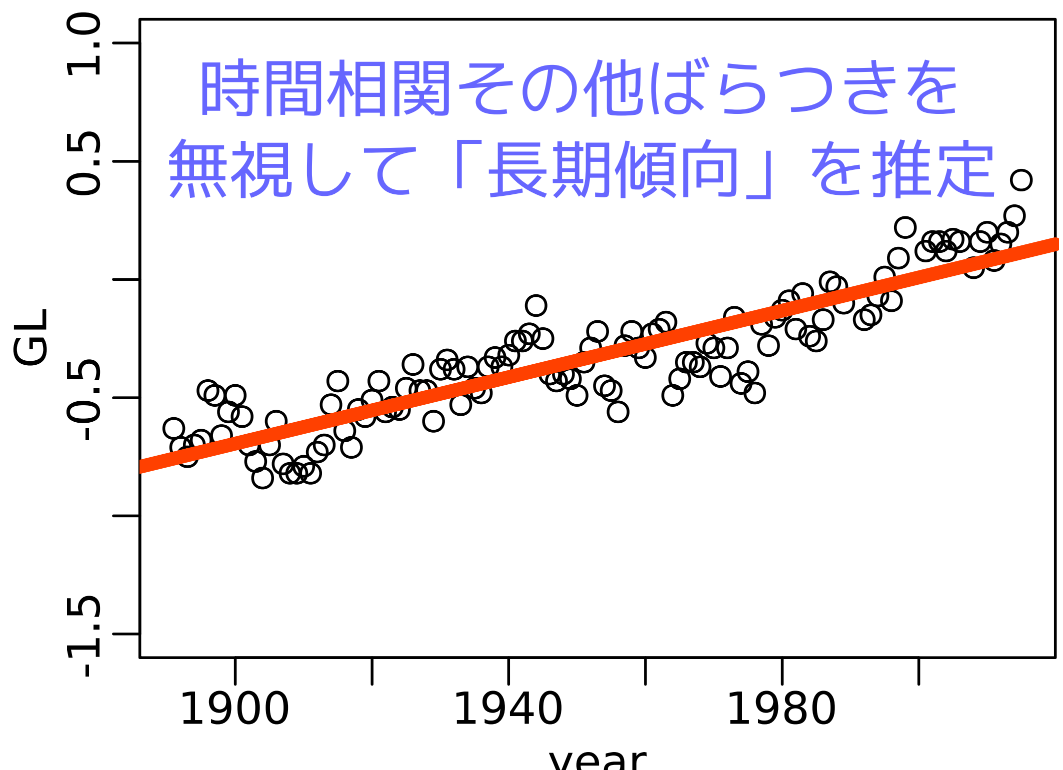


「とりあえず，直線回帰」の危険性

```
> 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℃

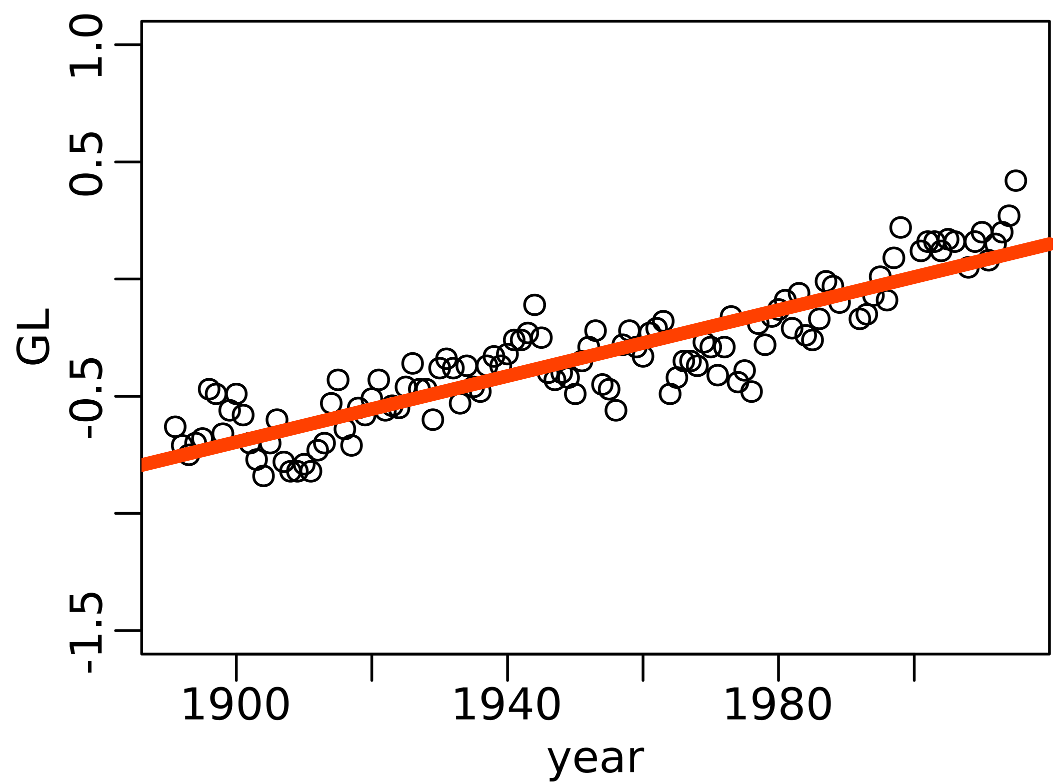
確率 1京ぶんの 2?



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
```

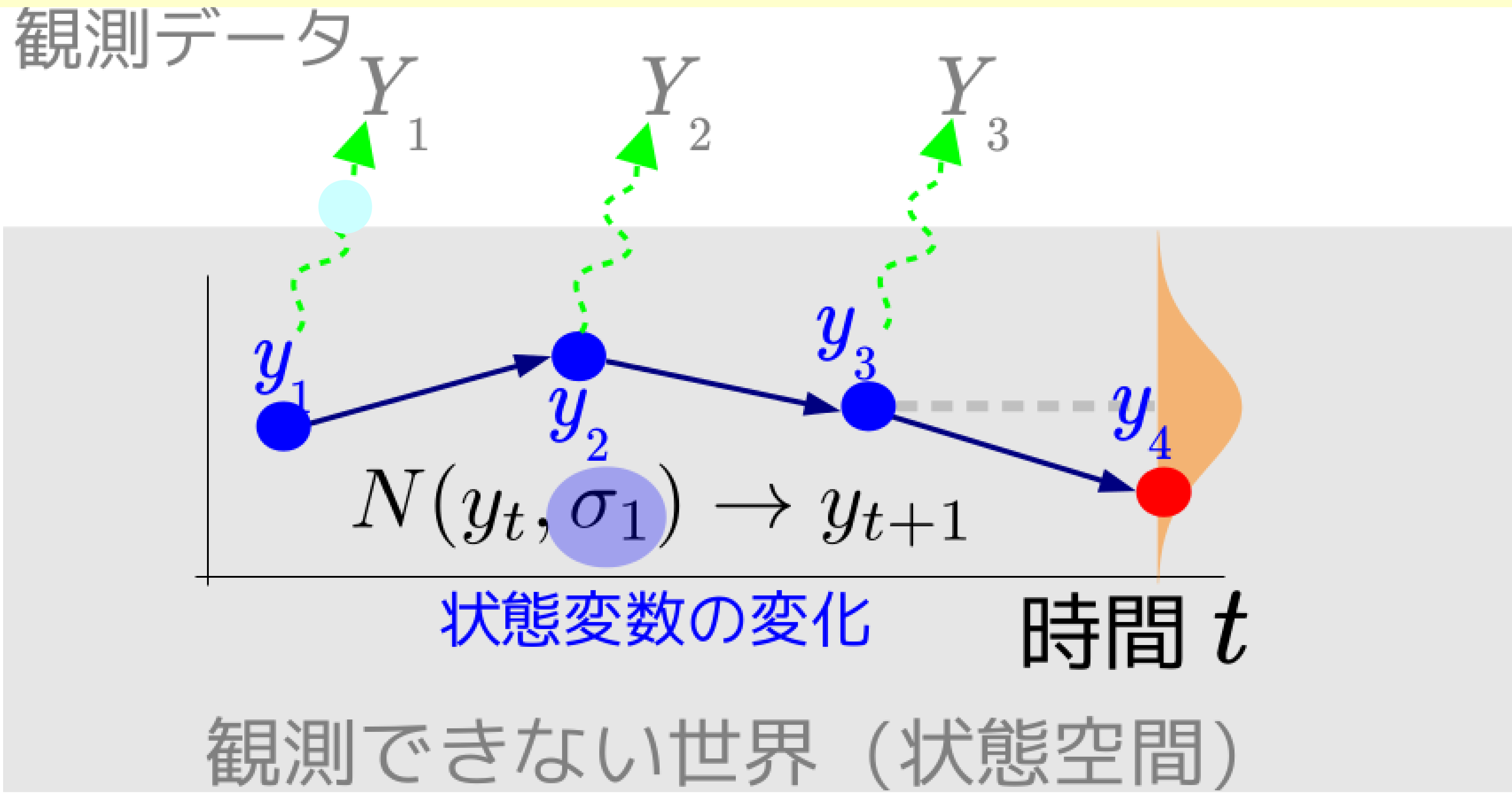
100年あたり 0.70℃



状態空間モデル：すべてを同時に推定

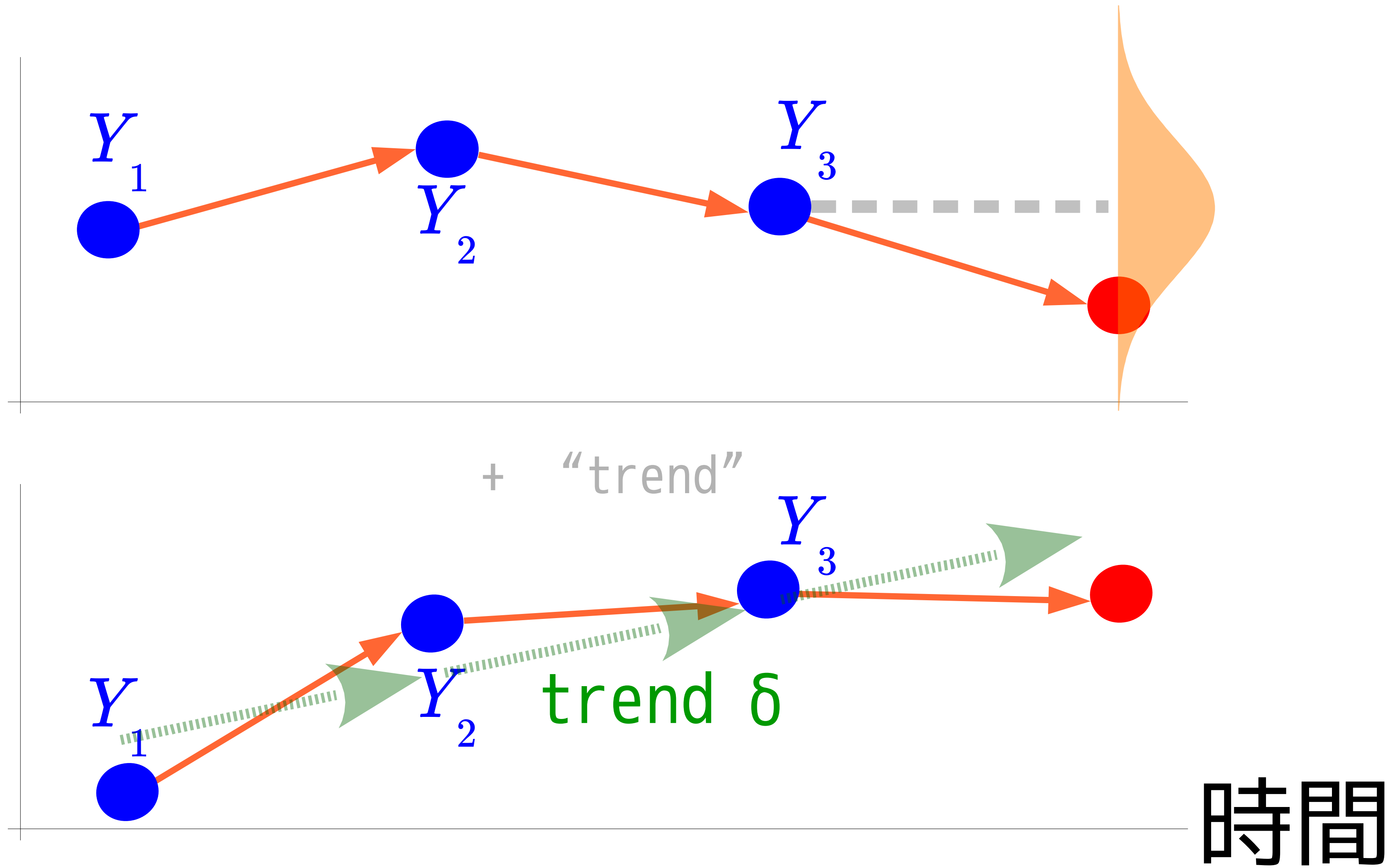
Hierarchical Bayesian state-space model

ランダムウォーク+各年独立なノイズ



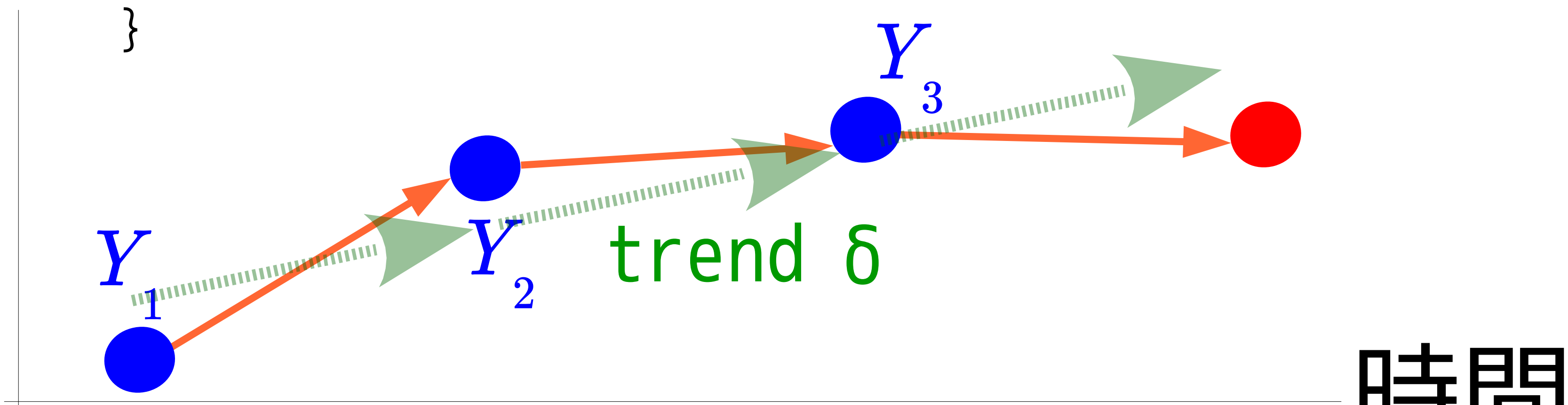
状態空間モデル：すべてを同時に推定

ランダムウォーク+各年独立なノイズ



状態空間モデル：すべてを同時に推定

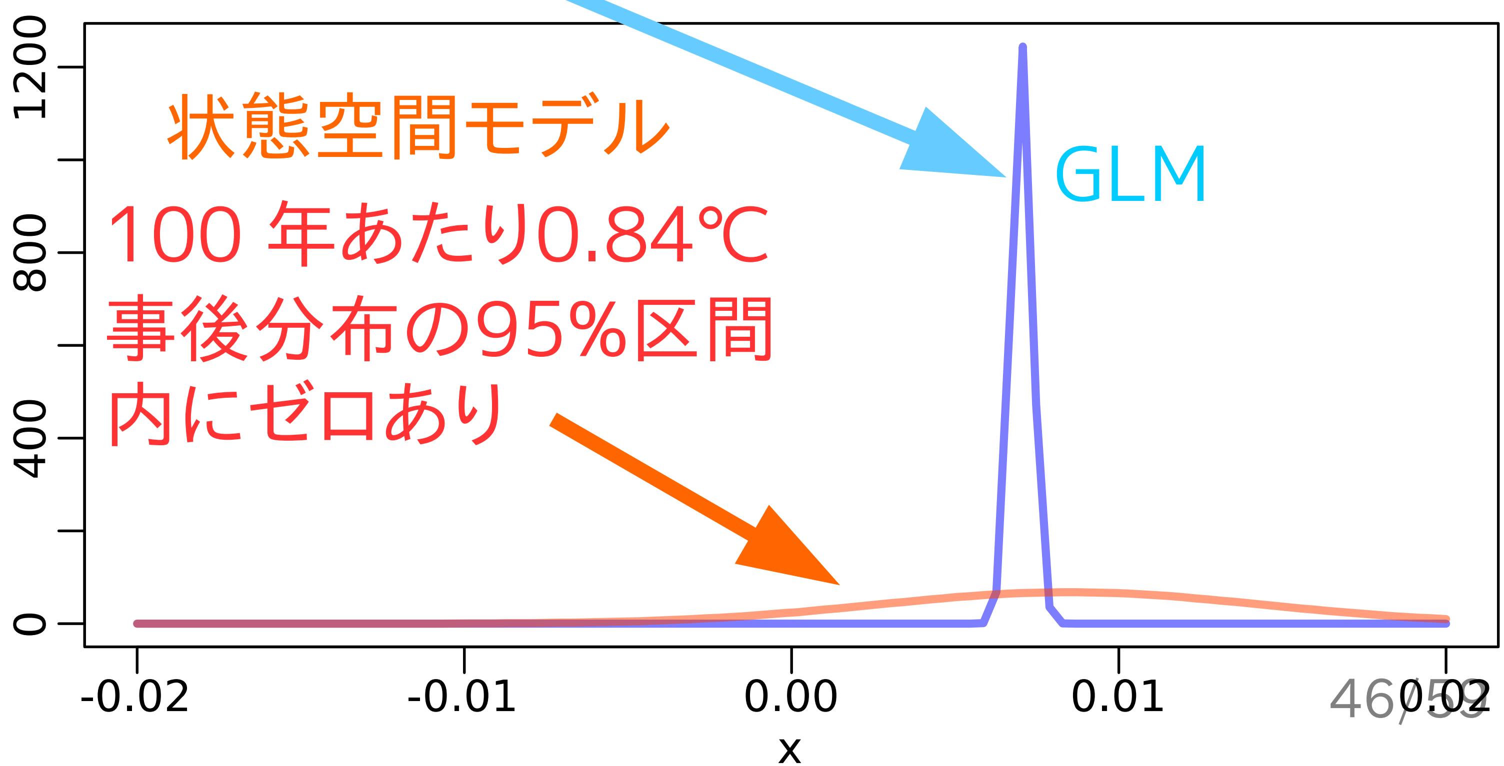
```
Y[1] ~ dnorm(y[1], tau[2])
y[1] ~ dnorm(0.0, Tau.Noninformative)
for (t in 2:N.Y) {
    Y[t] ~ dnorm(y[t], tau[2])
    y[t] ~ dnorm(m[t], tau[1])
    m[t] <- delta + y[t - 1]
}
delta ~ dnorm(0, Tau.Noninformative)
for (k in 1:2) {
    tau[k] <- 1.0 / (s[k] * s[k])
    s[k] ~ dunif(0, 1.0E+4)
}
```



状態空間モデルが予測した「温暖化」

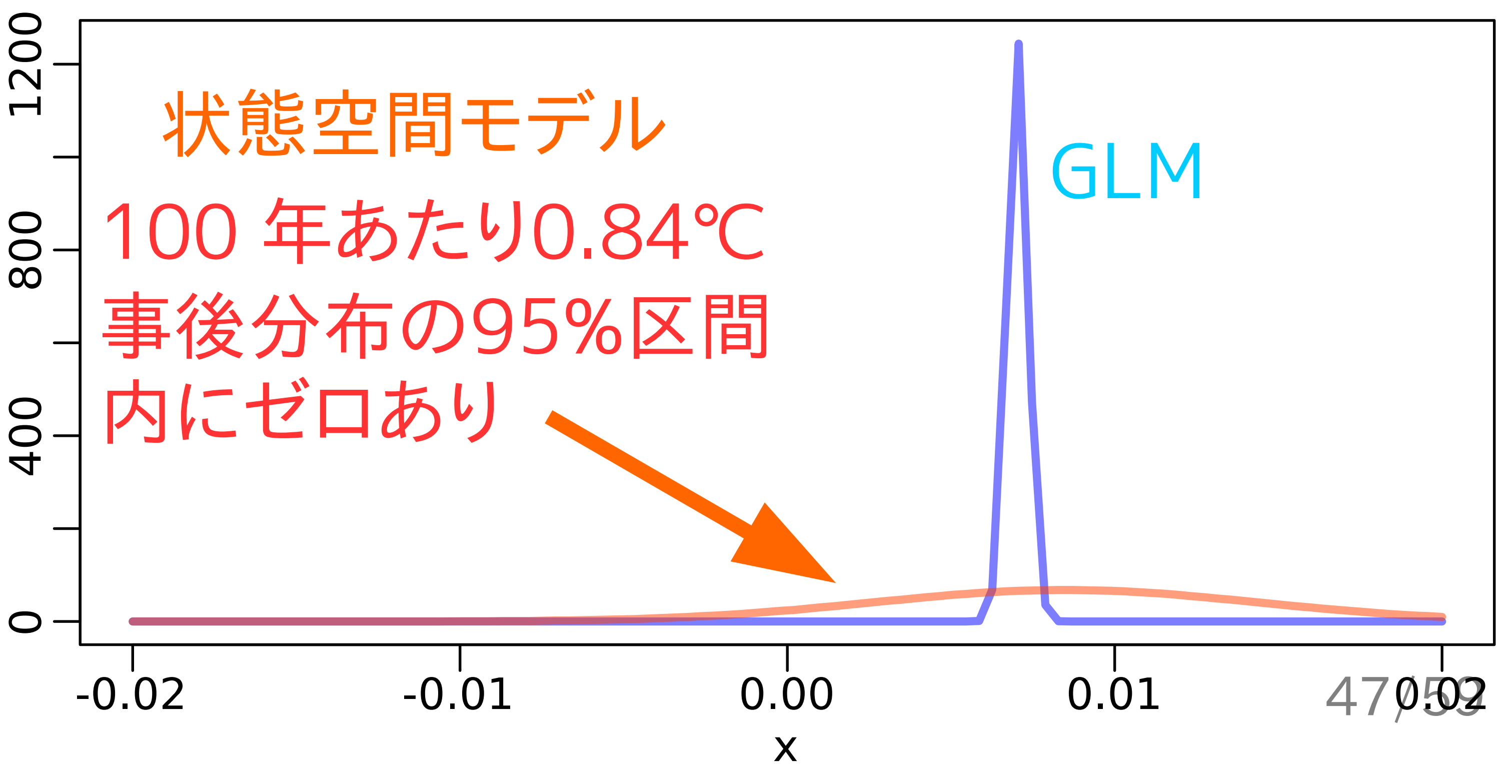
```
> summary(glm(GL ~ year, data = d))

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℃



観測値間に相関あり→サンプルサイズが小さくなる

100年
あたり
0.70℃



疑わしい回帰
spurious regression

時系列どうしの回帰

time series Y ~ time series X

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

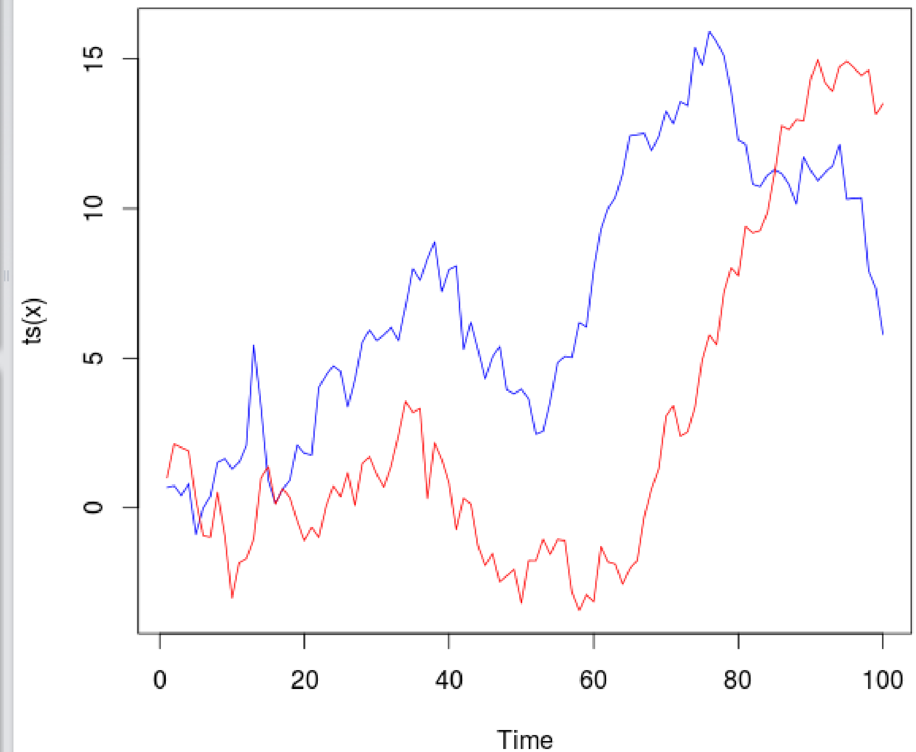
「見せかけの回帰」spurious regression

spurious_regression.R

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

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
```

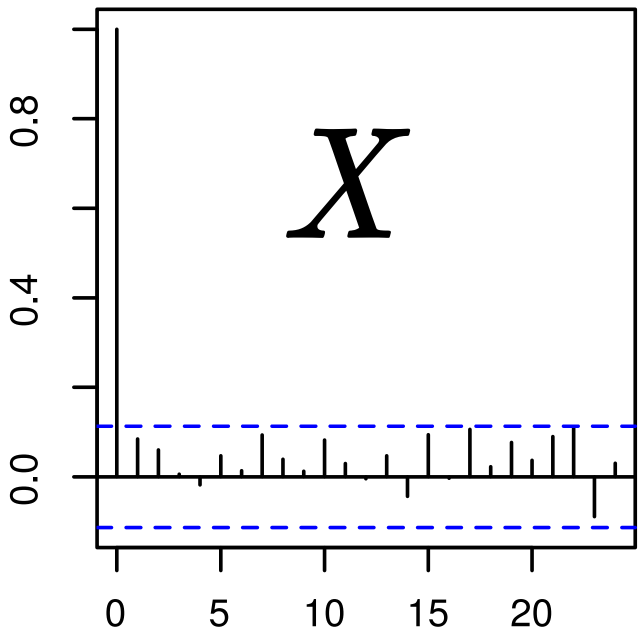


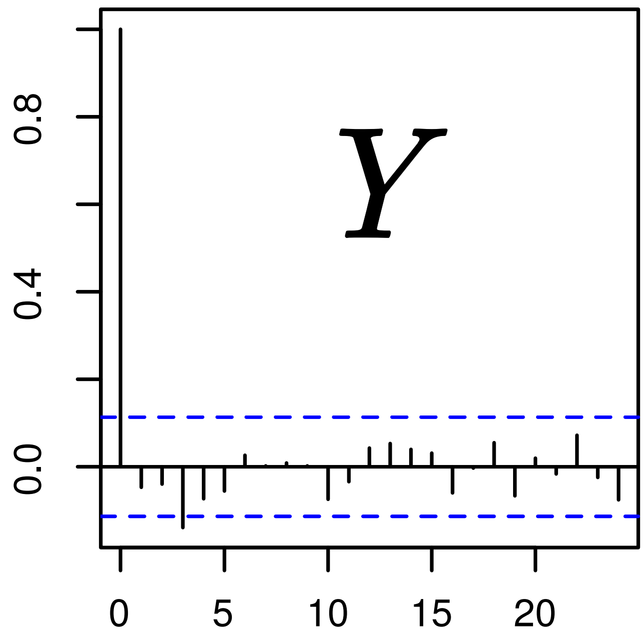
y_t $\quad$ x_t

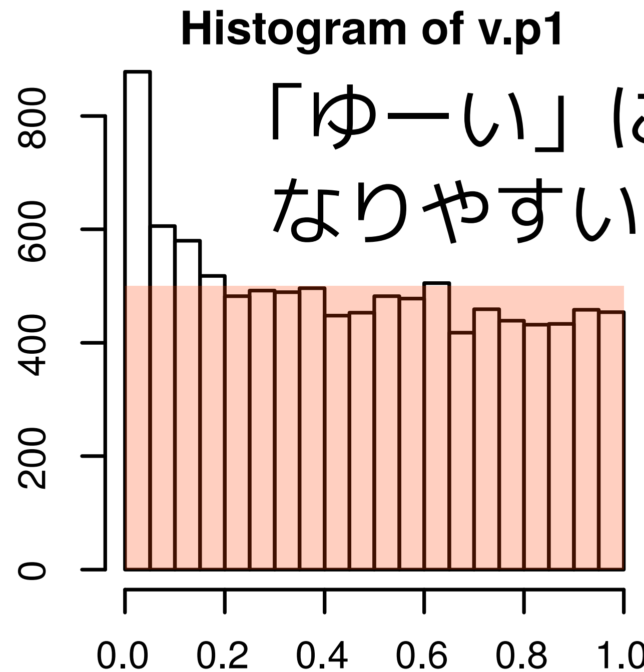
Time_series1 ~ Time_series2

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？







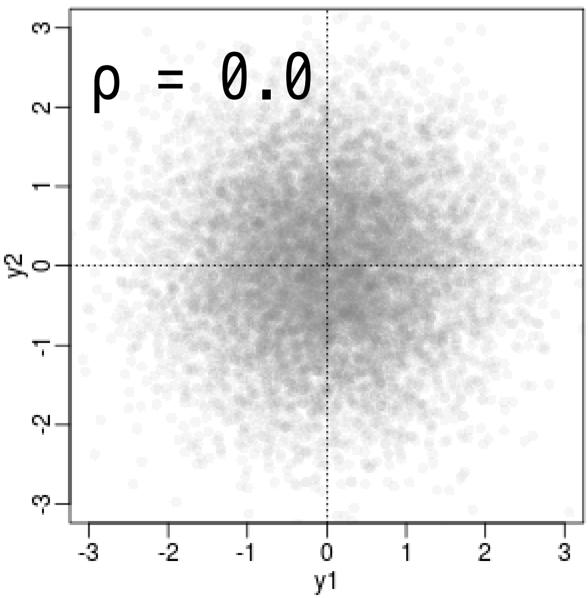
しかし glm(Y ~ X) とすると…

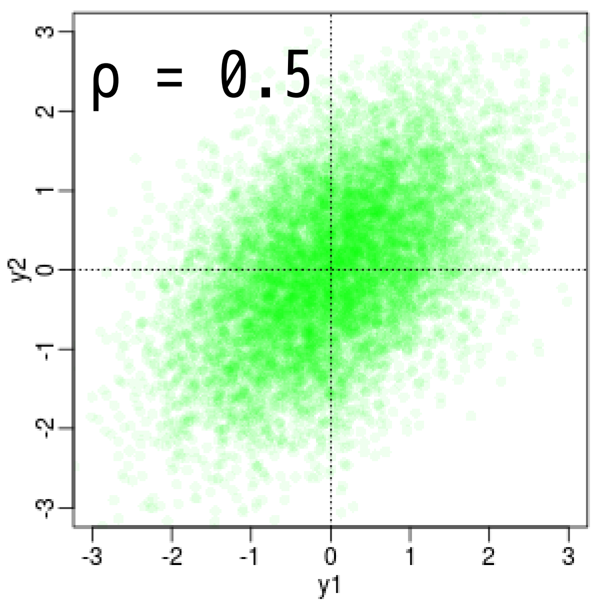
疑わしい回帰
spurious regression

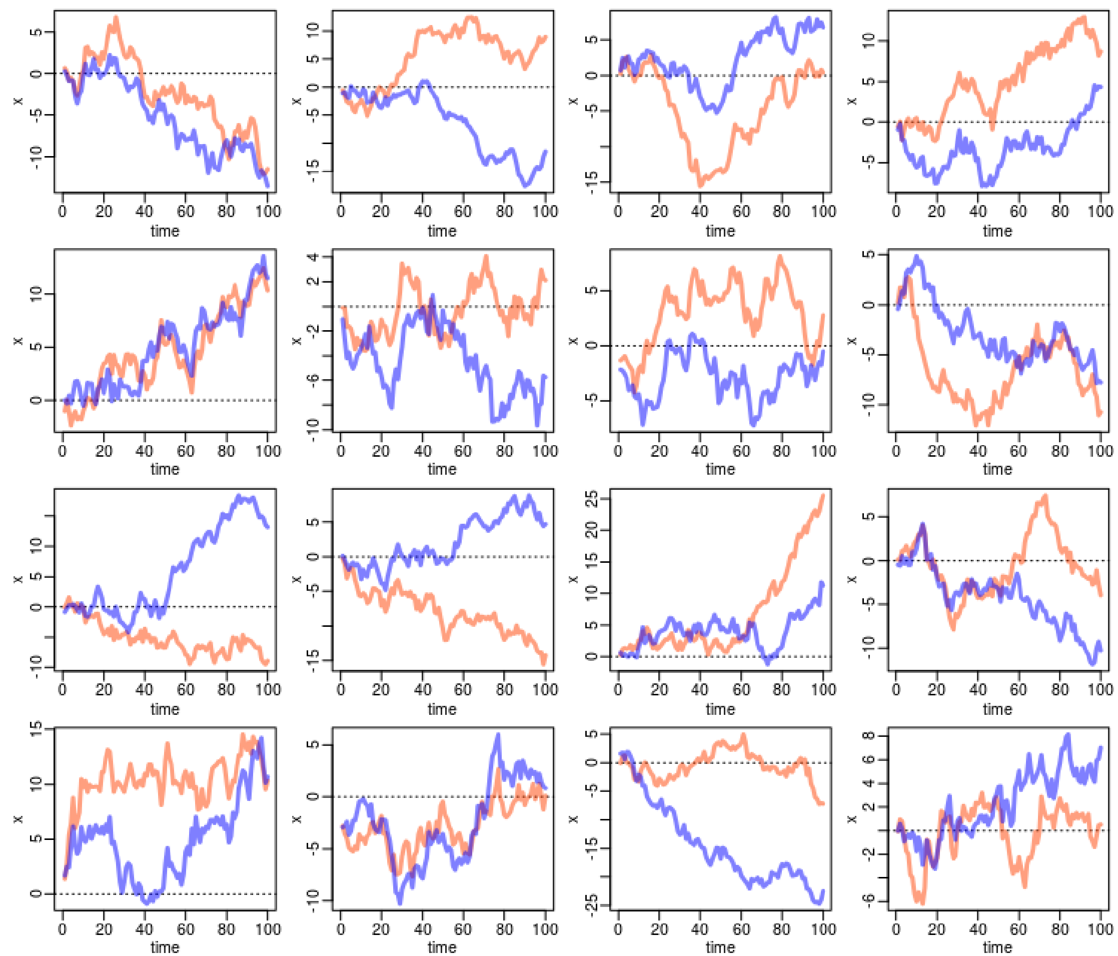
状態空間モデル (SSM)で

あつかえないか?

二変量正規分布とランダムウォーク







二変量正規分布を部品とする状態空間モデル

```
for (i in 1:N.Y) {
    Y[i, 1:2] ~ dmnorm(mu[1:2], Omega[1:2, 1:2])
}
mu[1] ~ dunif(-1.0E+4, 1.0E+4)
mu[2] ~ dunif(-1.0E+4, 1.0E+4)
Omega[1:2, 1:2] <- inverse(VarCov[1:2, 1:2])
VarCov[1, 1] <- sigma[1] * sigma[1]
VarCov[1, 2] <- sigma[1] * sigma[2] * rho
VarCov[2, 1] <- sigma[2] * sigma[1] * rho
VarCov[2, 2] <- sigma[2] * sigma[2]
sigma[1] ~ dunif(0.0, 1.0E+4)
sigma[2] ~ dunif(0.0, 1.0E+4)
rho ~ dunif(-1.0, 1.0)
```

(R で実演)

階層ベイズモデルである
状態空間モデル
から得られた事後分布

```
 3 chains, each with 5200 iterations (first 200 discarded)
 n.sims = 15000 iterations saved
           mean    sd   2.5%    25%    50%    75% 97.5%  Rhat n.eff
mu[1]    -0.122 0.110 -0.342 -0.195 -0.120 -0.048 0.090 1.001  6000
mu[2]    -0.157 0.100 -0.355 -0.224 -0.157 -0.091 0.041 1.002  1500
sigma[1]  1.091 0.079  0.949  1.036  1.086  1.142 1.261 1.001  6100
sigma[2]  0.993 0.074  0.864  0.941  0.987  1.039 1.151 1.001  4100
rho       0.568 0.070  0.420  0.523  0.573  0.617 0.693 1.001 11000
```

ふたつの時系列データの変動が
相関しているかどうかを特定できる

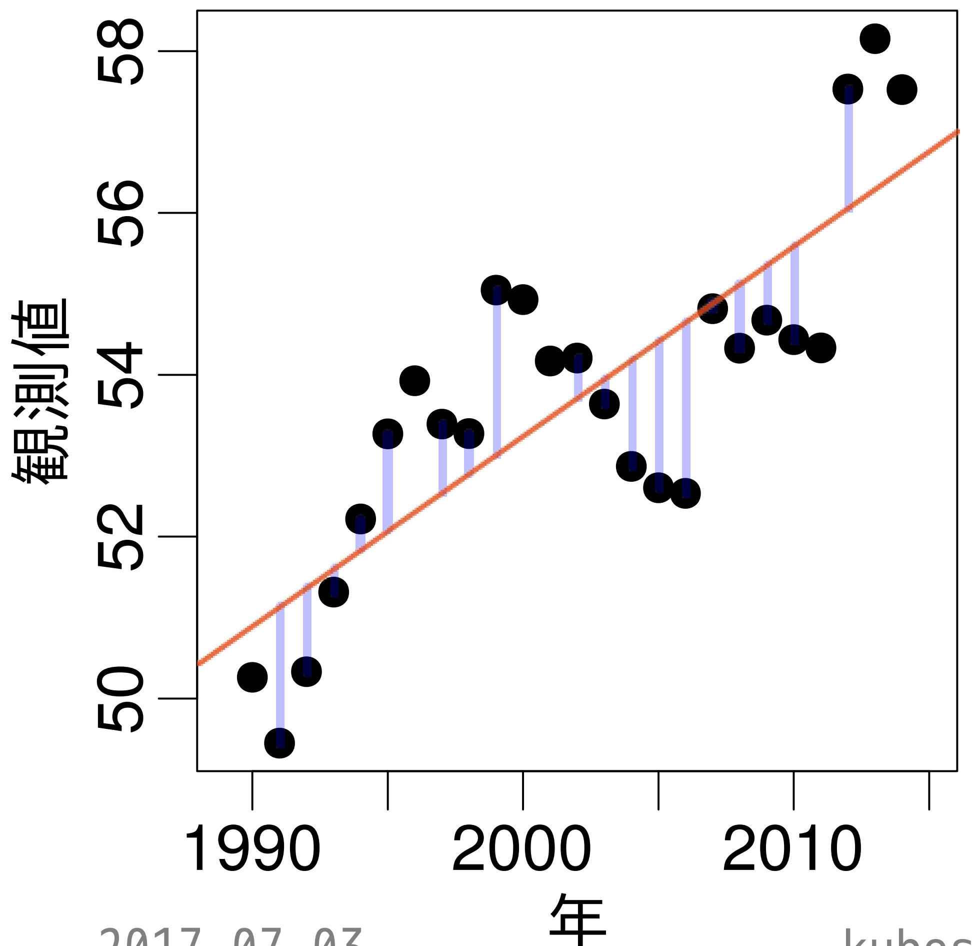
おわりに

時間的な相関はデータの

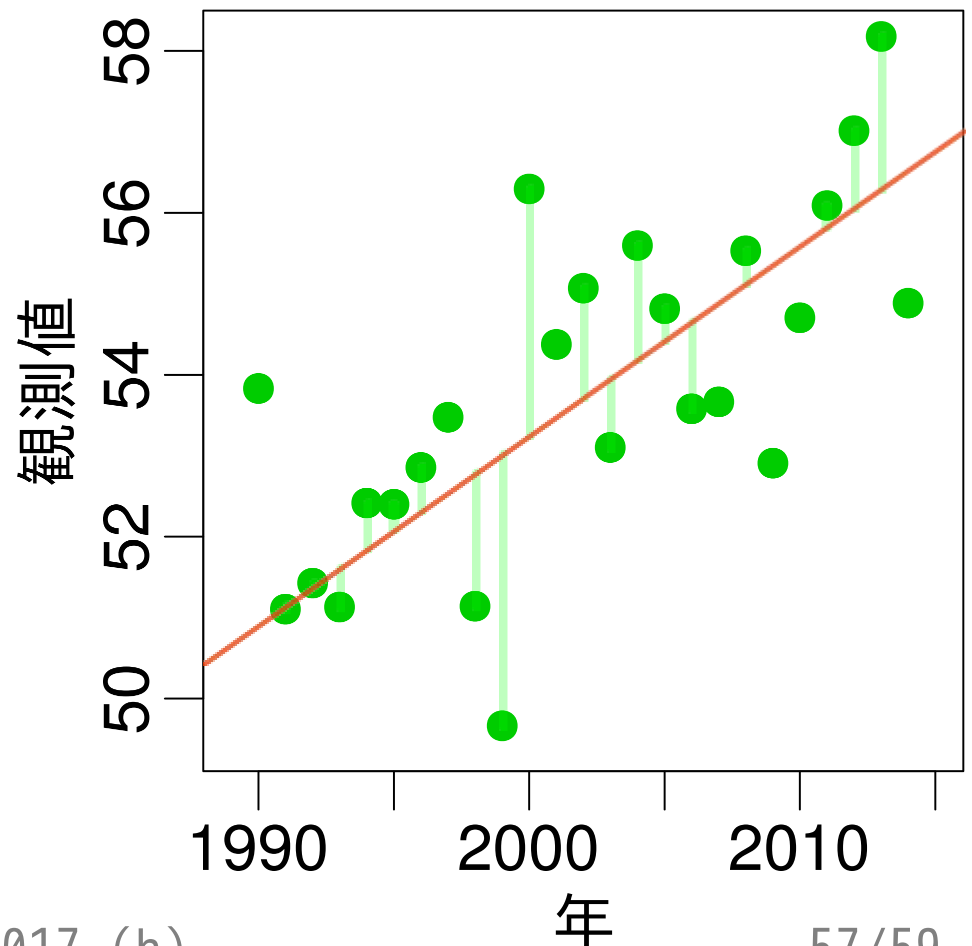
情報量を減少させる

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

おしまい

The Evolution of Linear Models

Parameter Estimation

MCMC

Hierarchical Bayesian Model
(HBM)

Generalized Linear Mixed Model
(GLMM)

MLE

Generalized Linear Model (GLM)

MSE

Linear Model

データ解析は

階層ベイズモデルで！