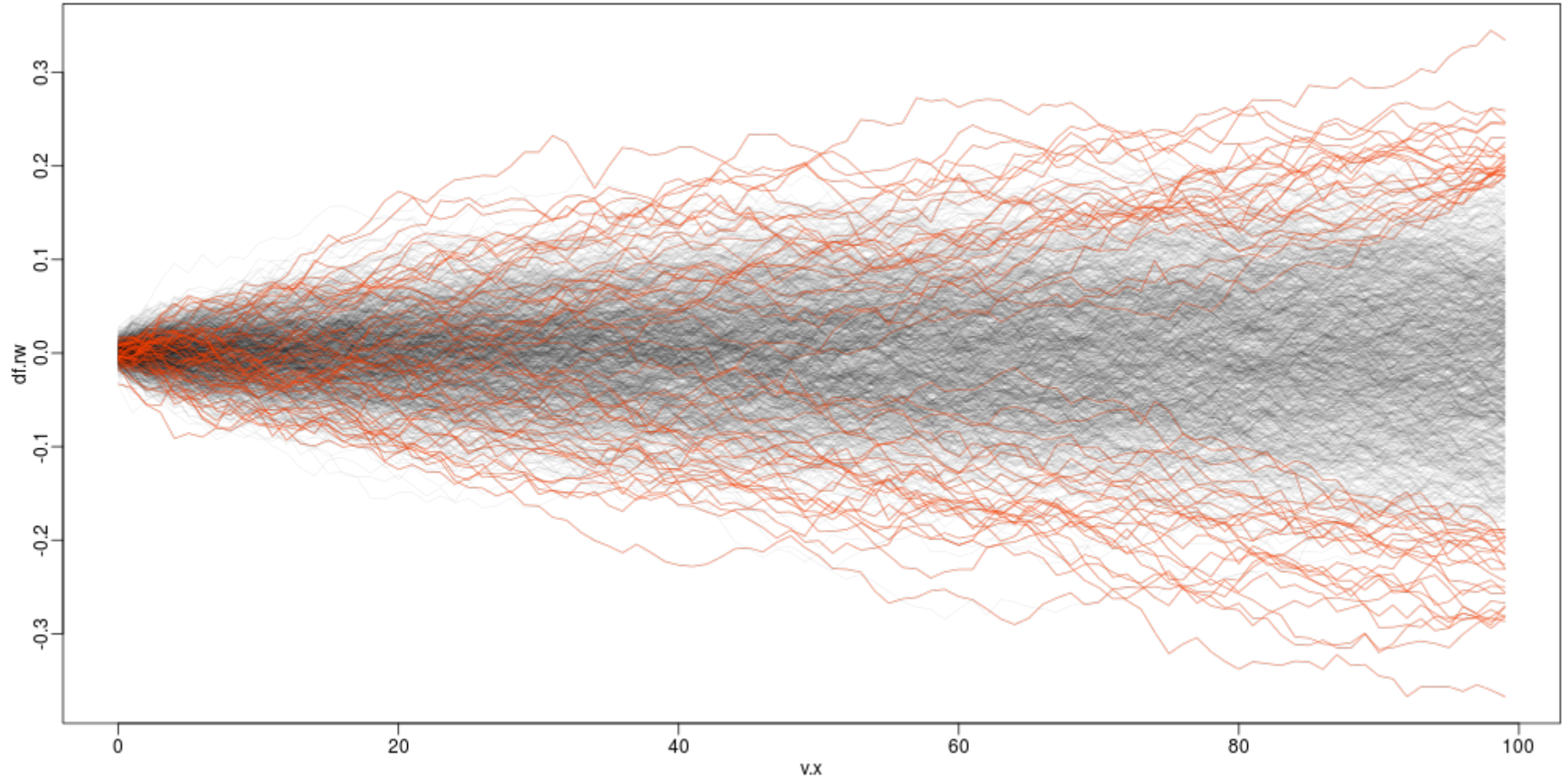


# 階層ベイズモデル - 時間変化の統計モデリング

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!

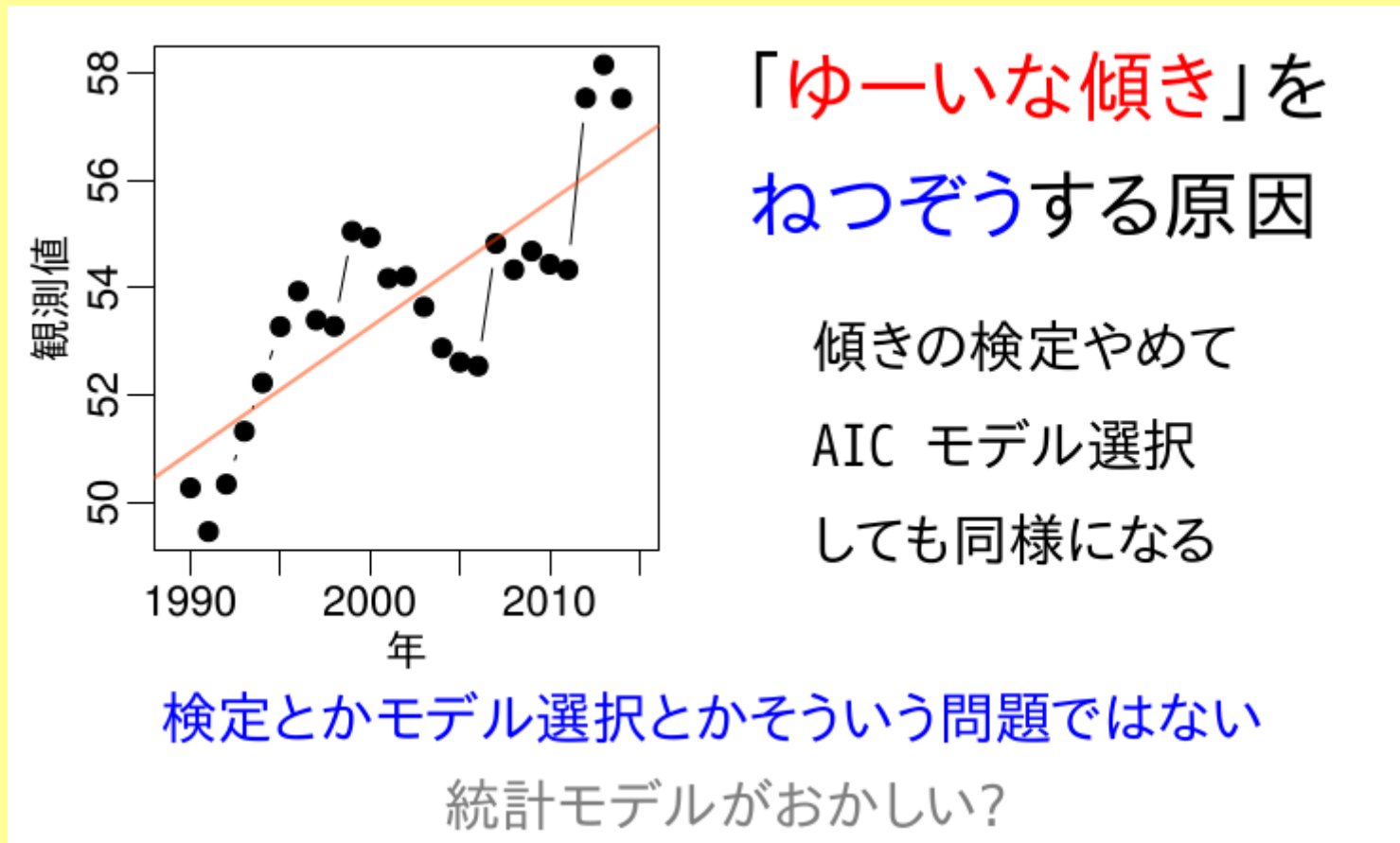
統計モデル  
のあてはめ

(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  $\sim$  気温 とか  
(これは次回)

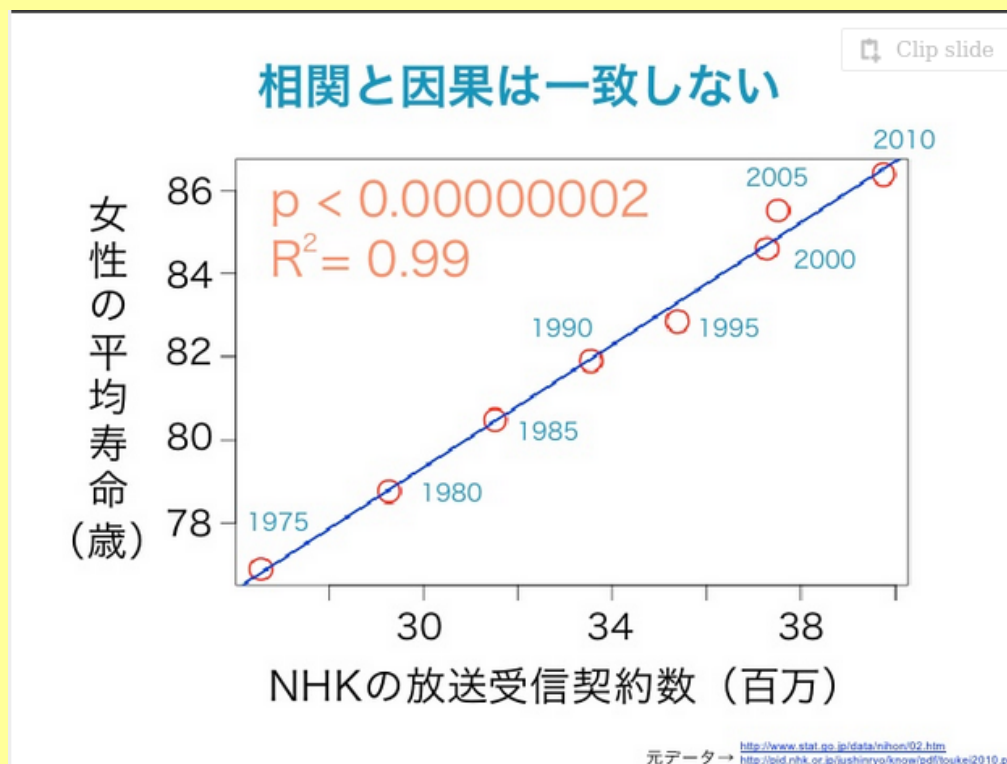
# (危1) 時系列データを GLM で



# (危2) 時系列 $Y_t \sim$ 時系列 $X_t$

「相関は因果関係ではない」

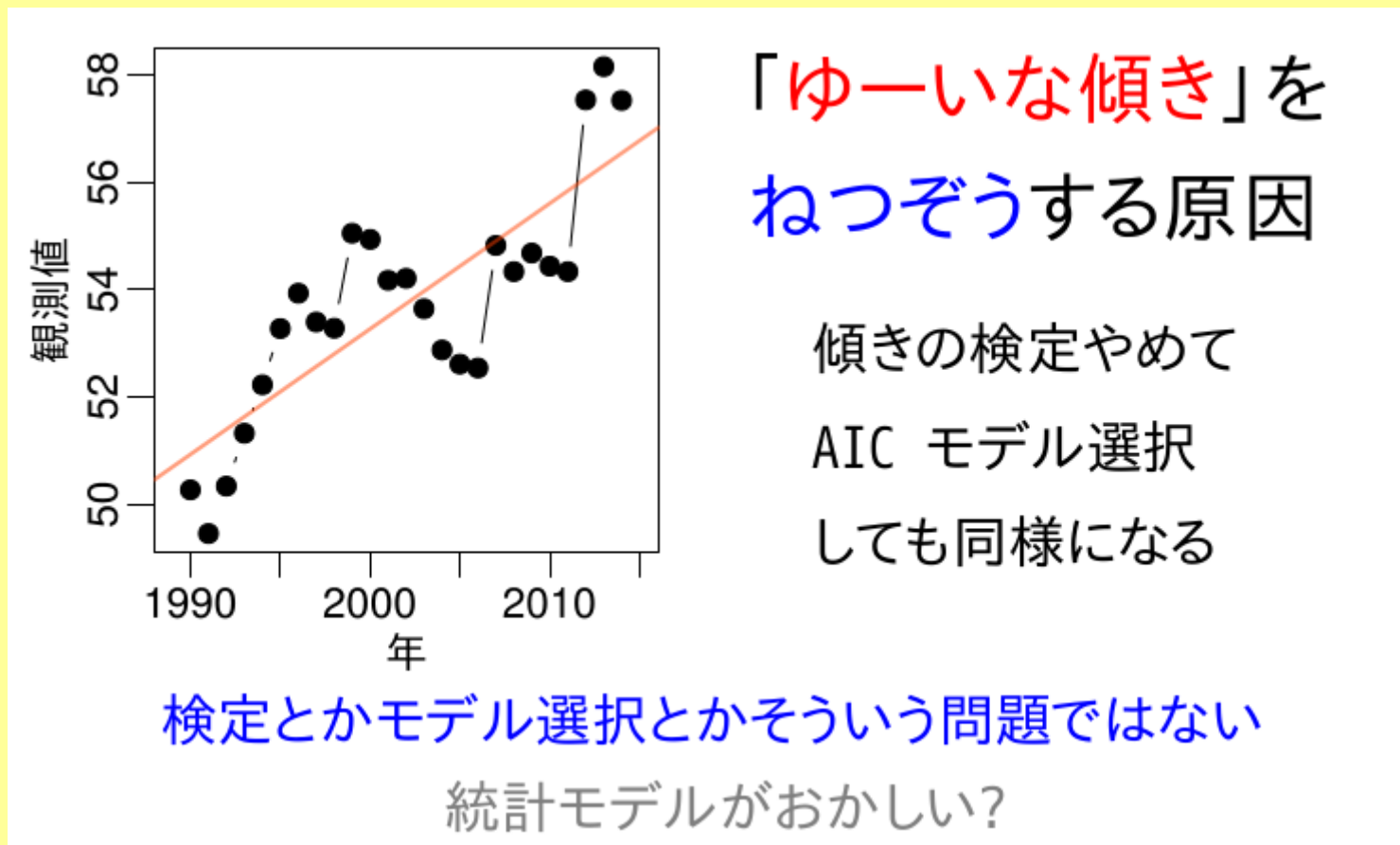
問題の一部：**にせの回帰** (これは次回)



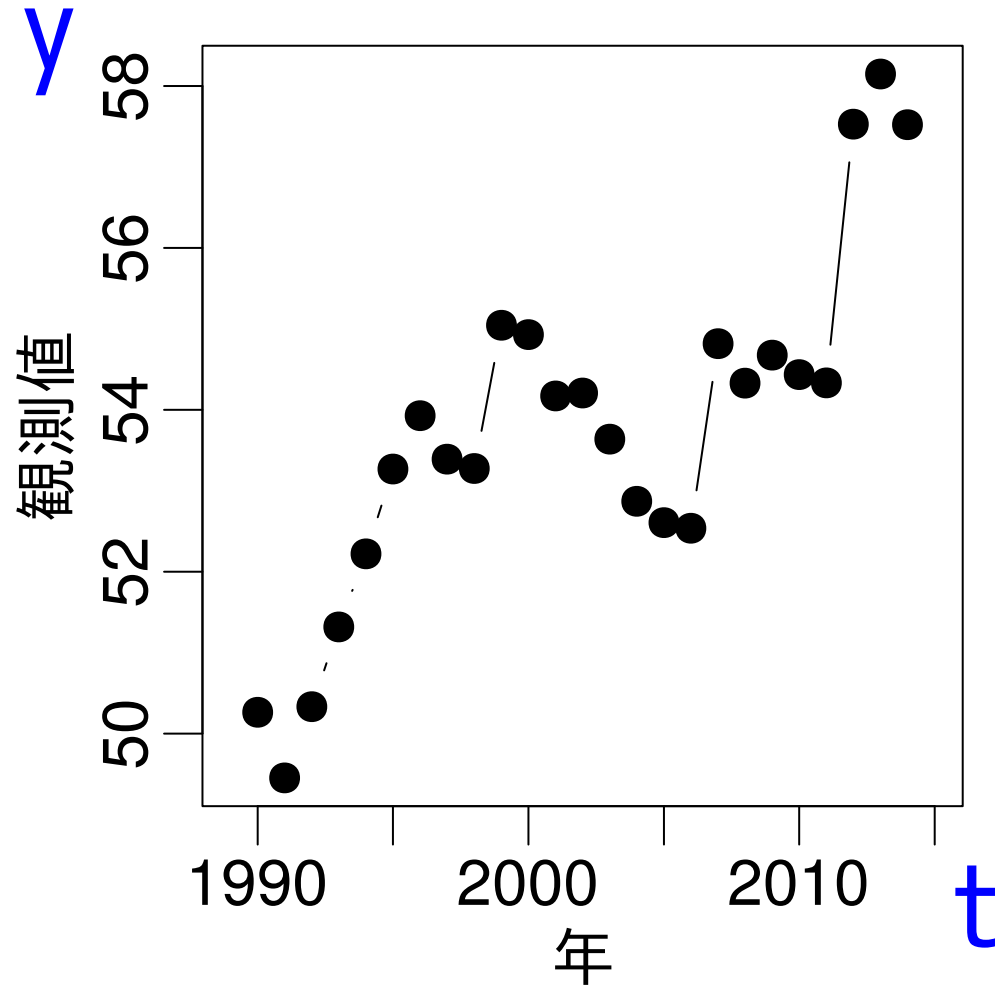
# 時系列データの統計モデリング

- 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



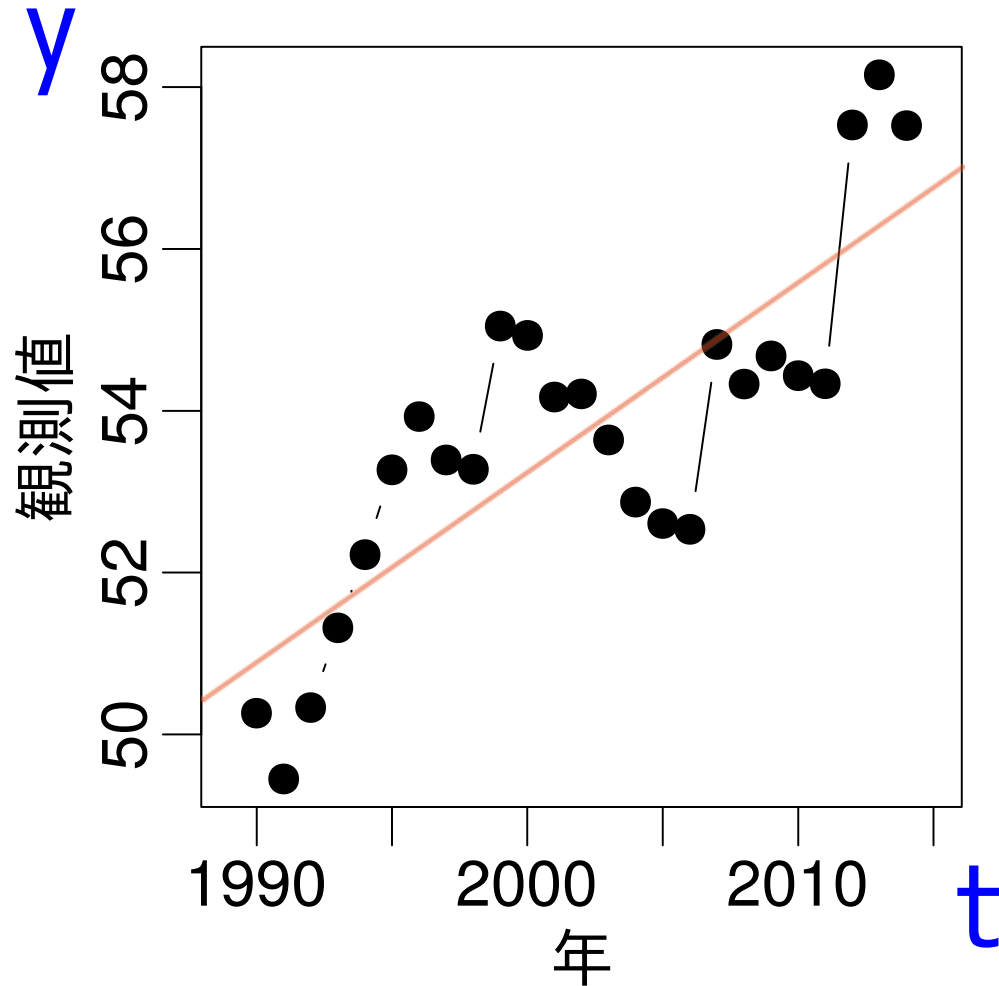
このような時系列データがあったとしましょう



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

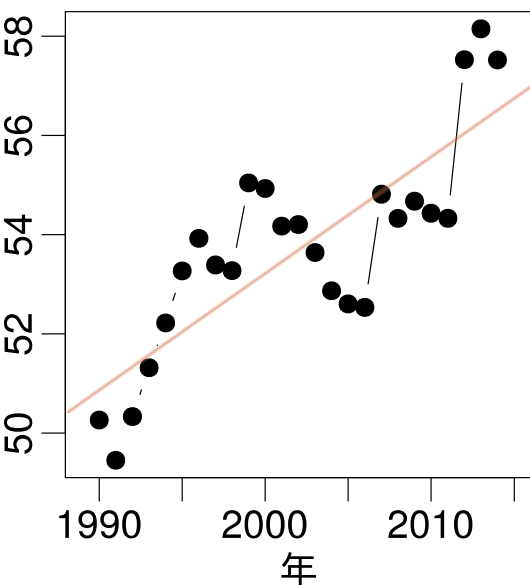


$\text{glm}(y \sim t)$

…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:

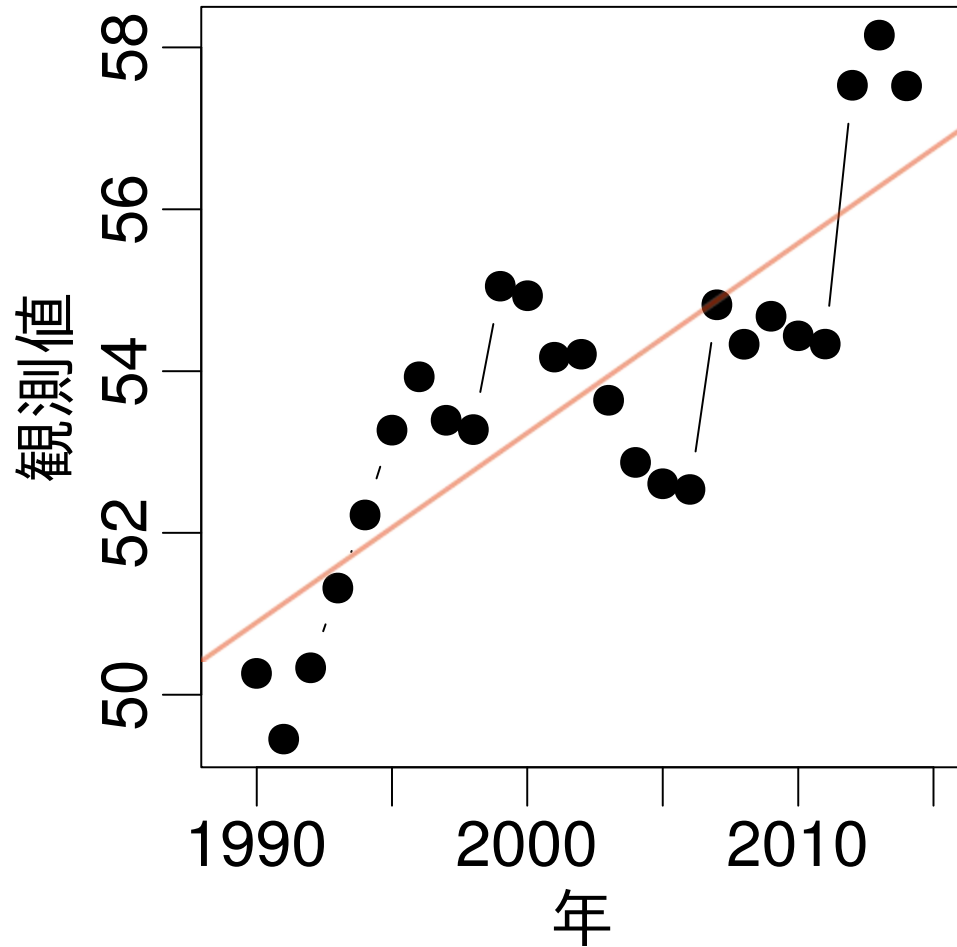
Min	1Q	Median	3Q	Max
-2.1295	-1.0583	-0.0817	0.9860	2.0188

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



「ゆるい傾き」(偽)

が「ぞろぞろ」でます

傾きの検定やめて

AIC モデル選択

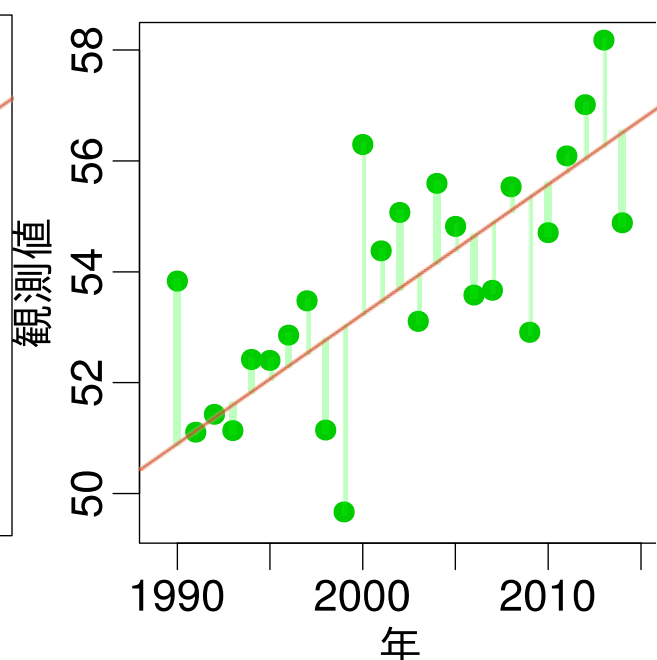
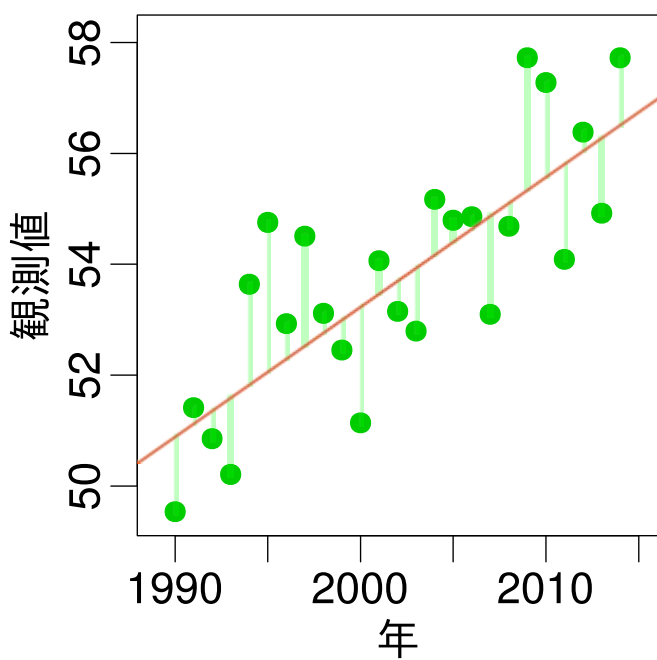
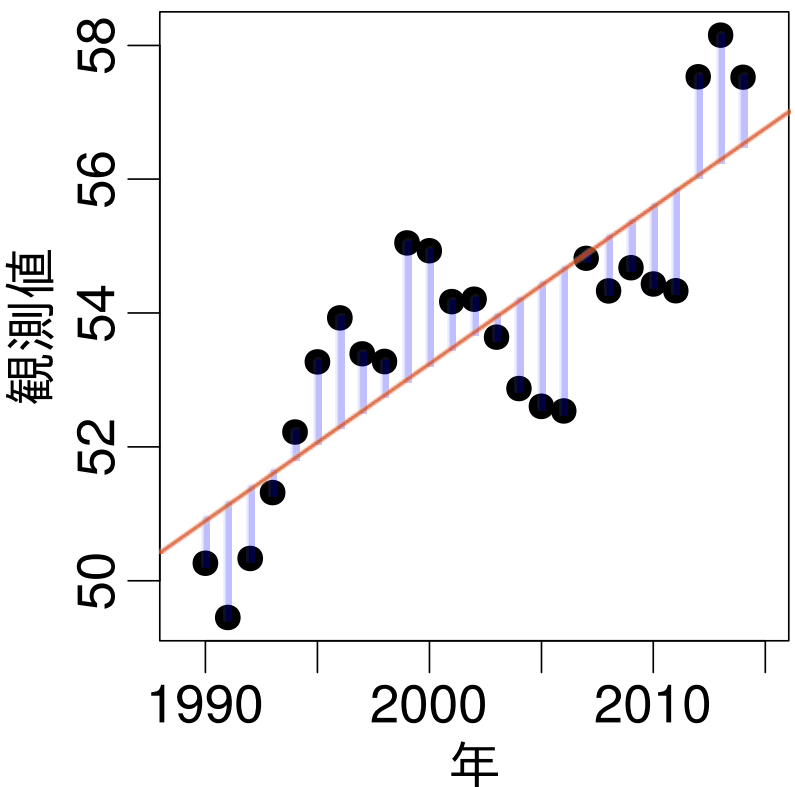
しても同様になる

検定とかモデル選択とかそういう問題ではない

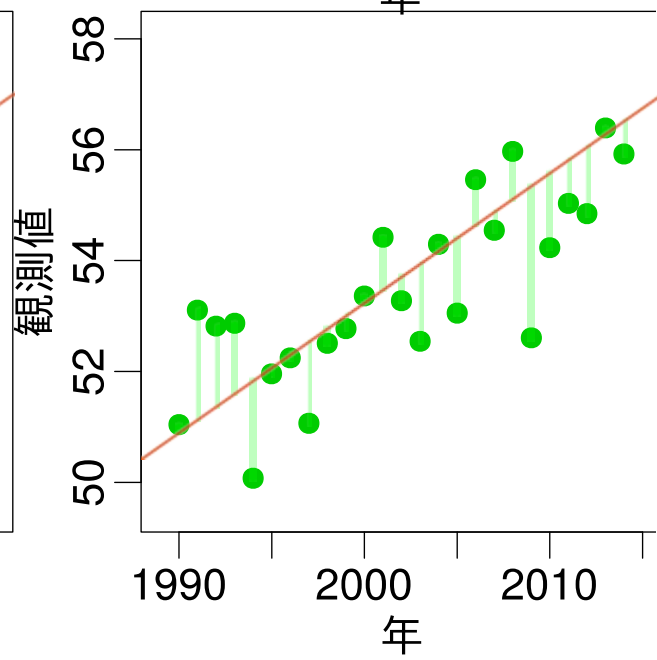
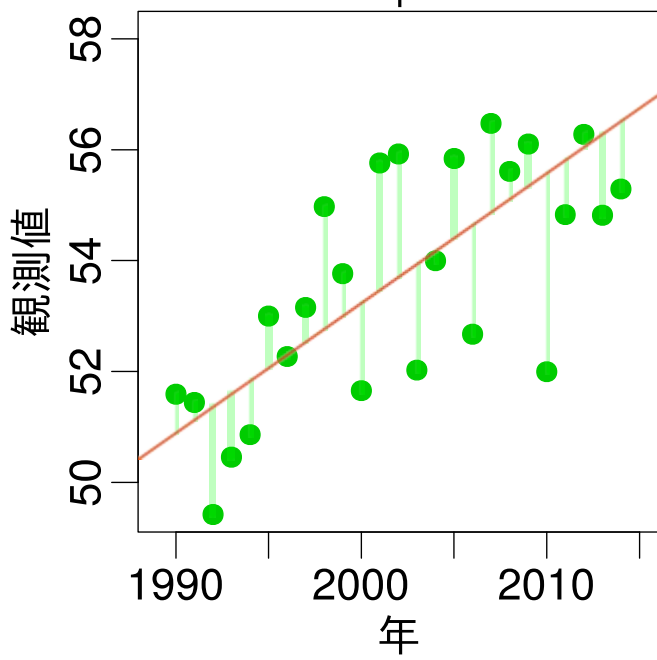
統計モデルがおかしい?

# 時系列の「ずれ」

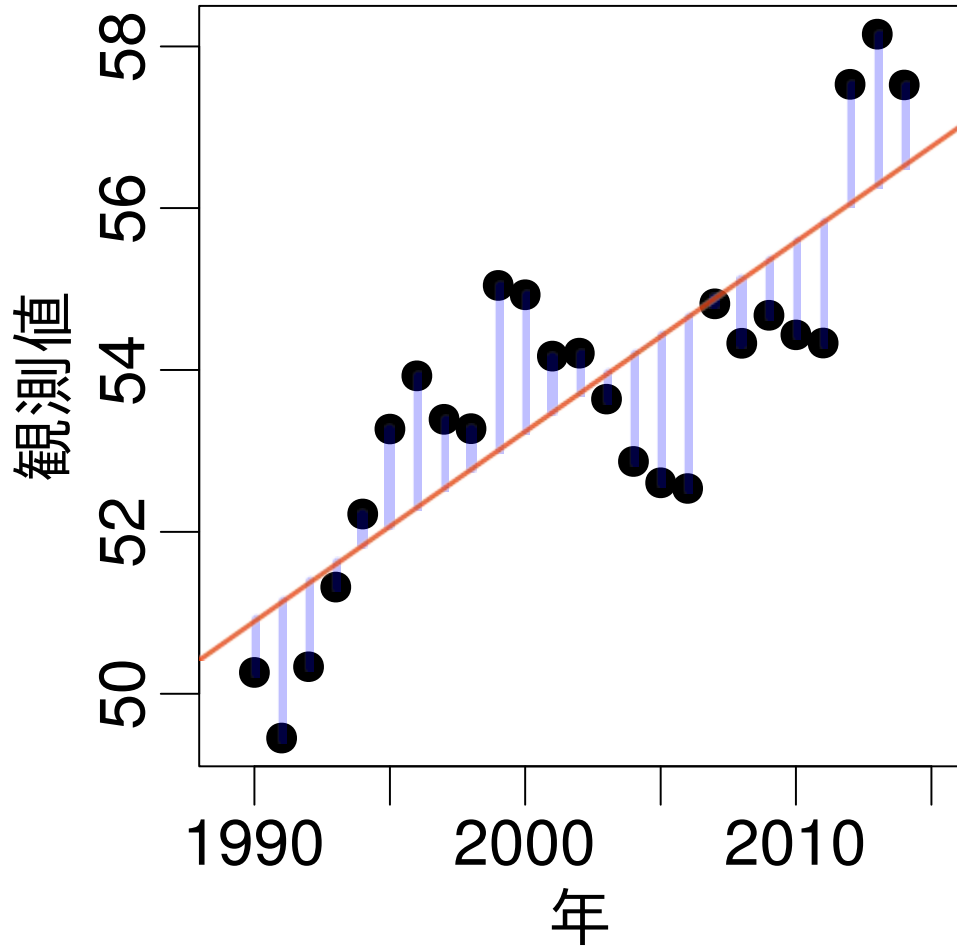
# GLM のずれ



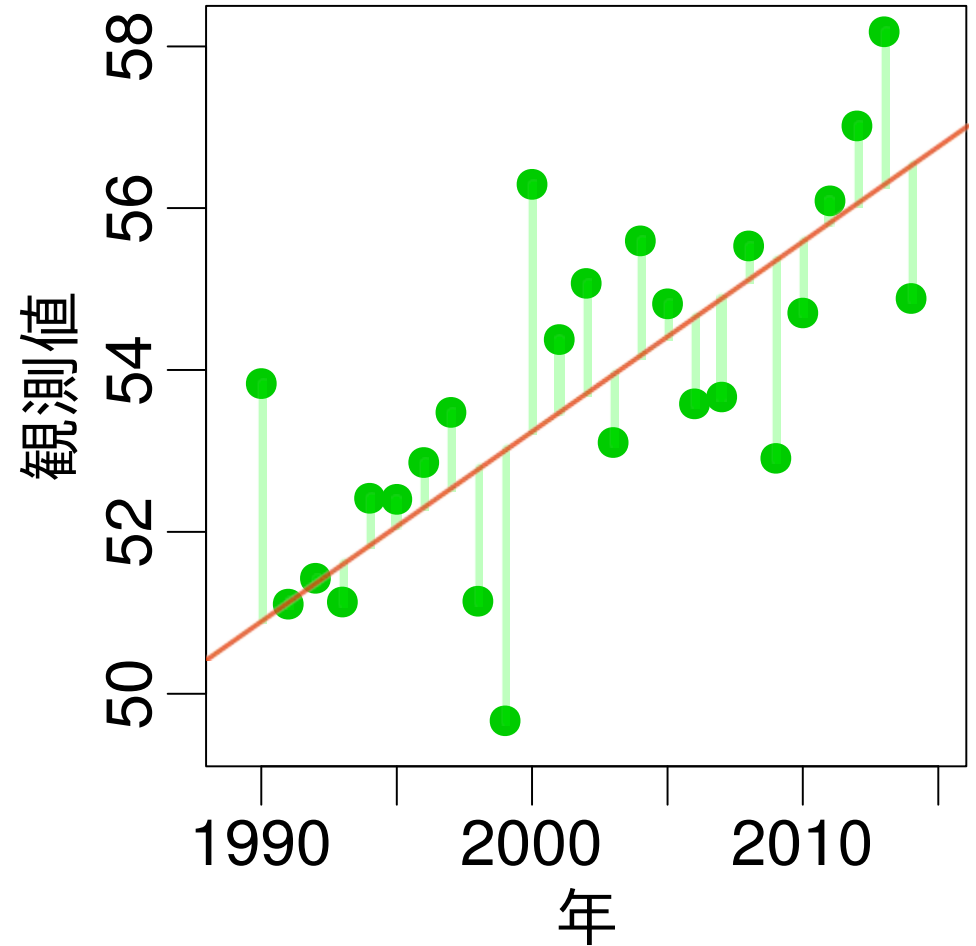
ずれかたが  
ちがってる?



# 時系列の「ずれ」



# GLM のずれ

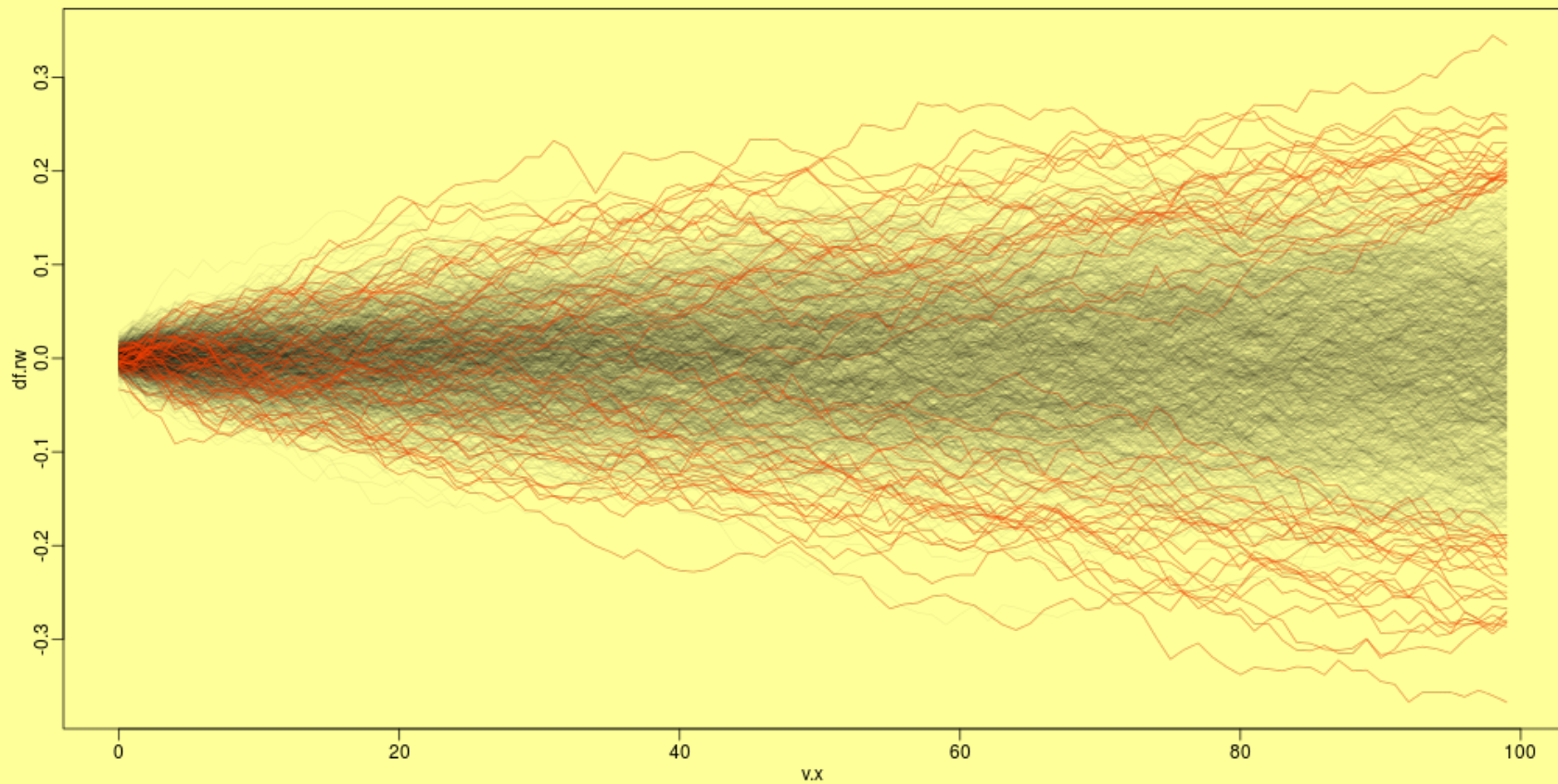


直線からのずれがちがう！

時間的自己相関がある

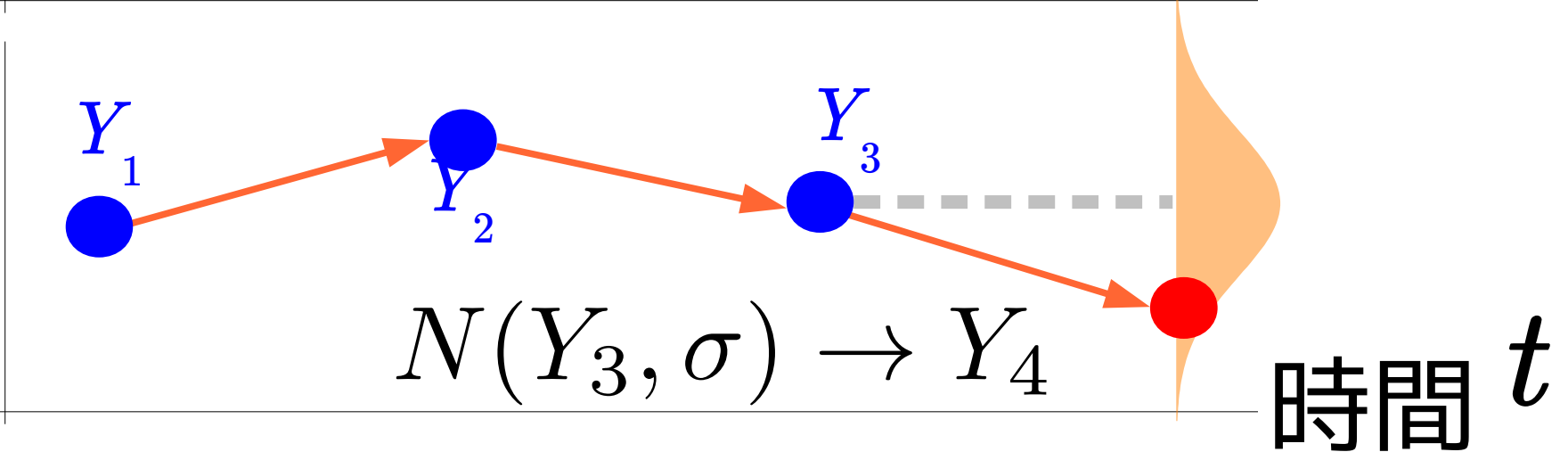
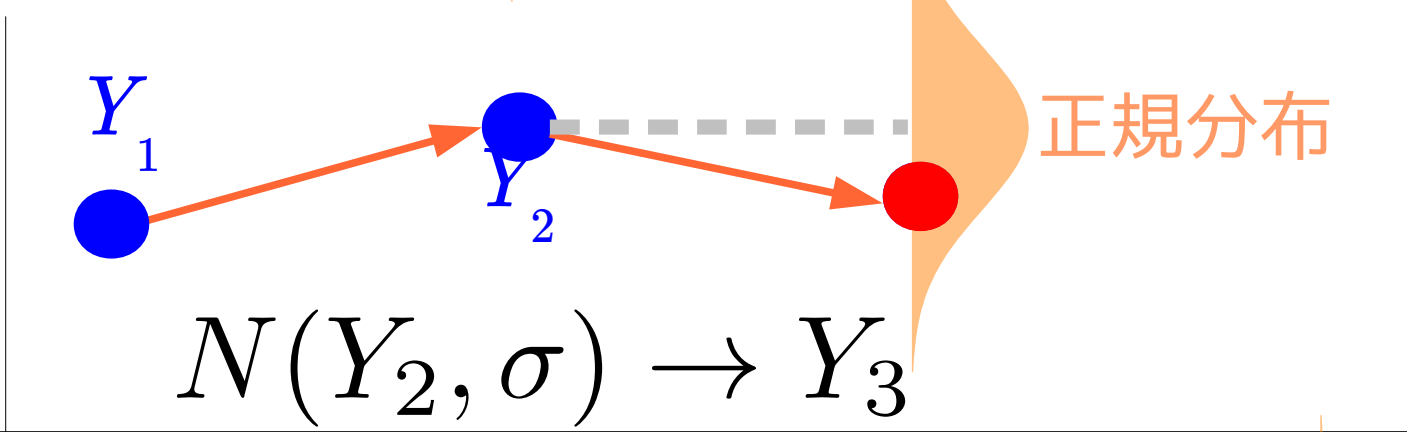
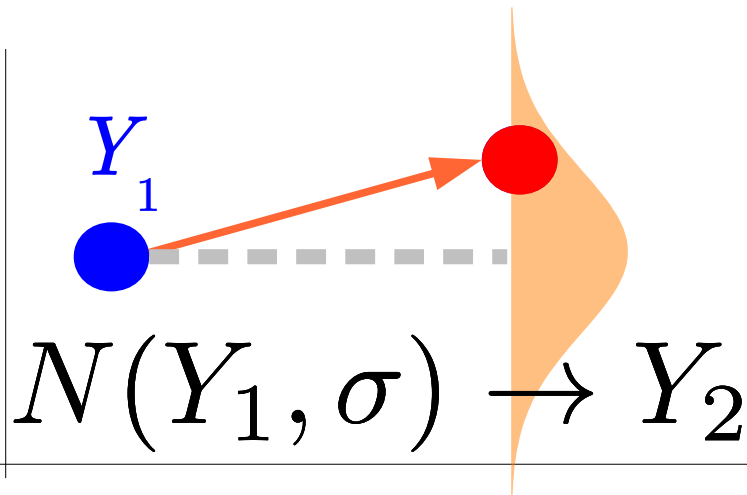
時間的自己相関がない

# 時系列の基本モデルのひとつ ランダムウォーク（乱歩）



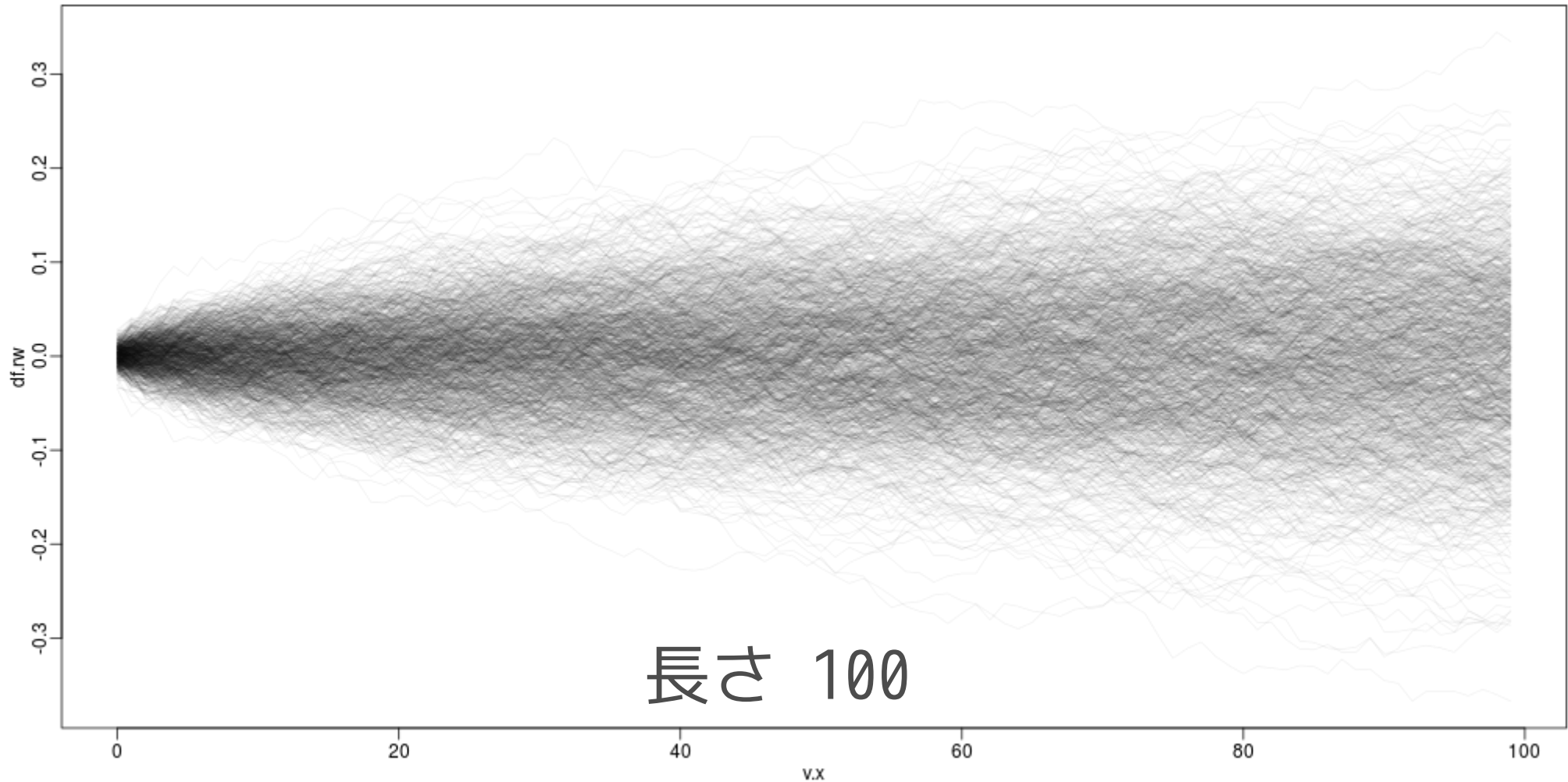
変数  
 $Y$

ランダムウォーク  
もっとも単純な  
モデル



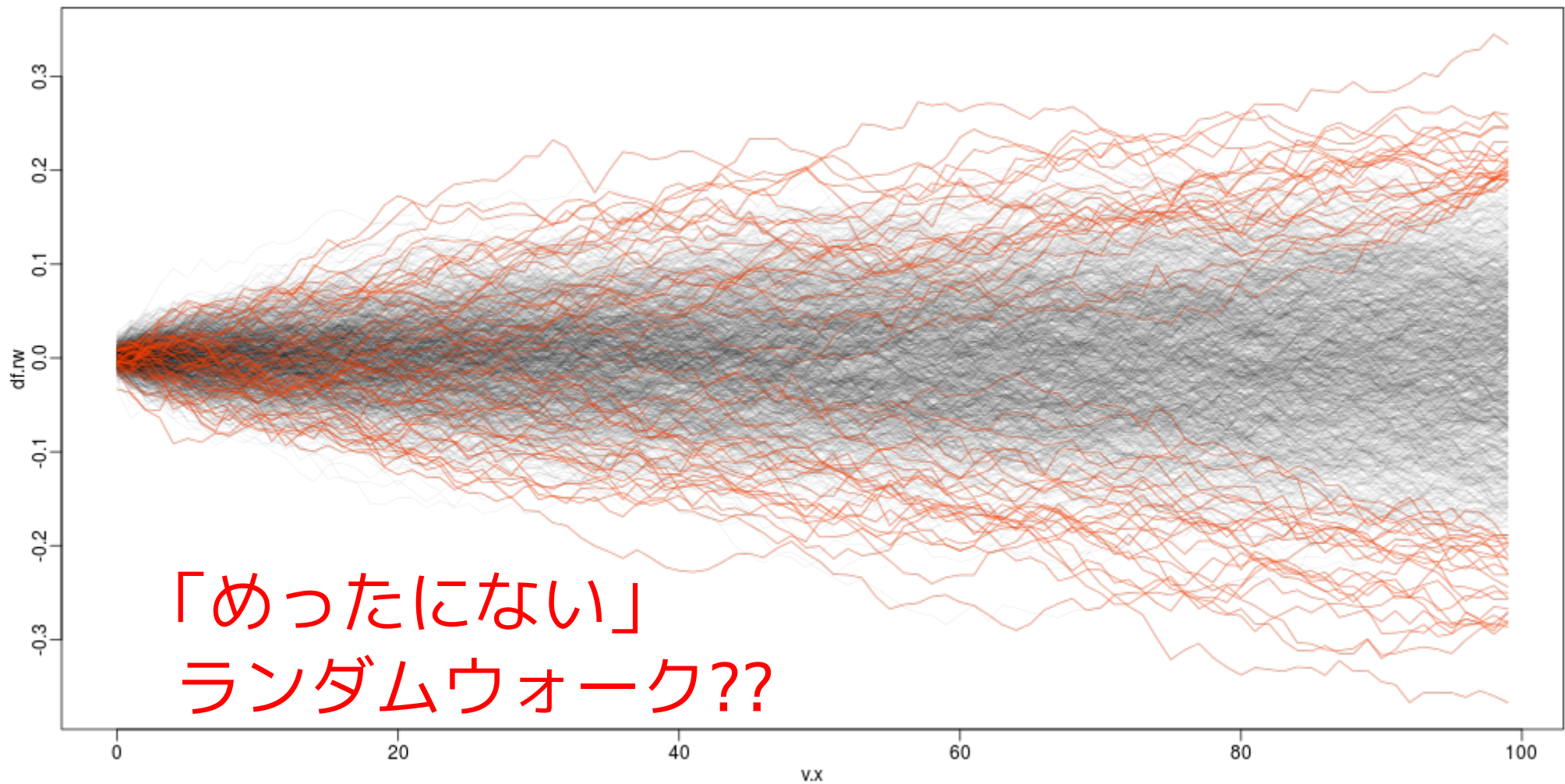
# ランダムウォークなサンプル時系列

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るのはありえる

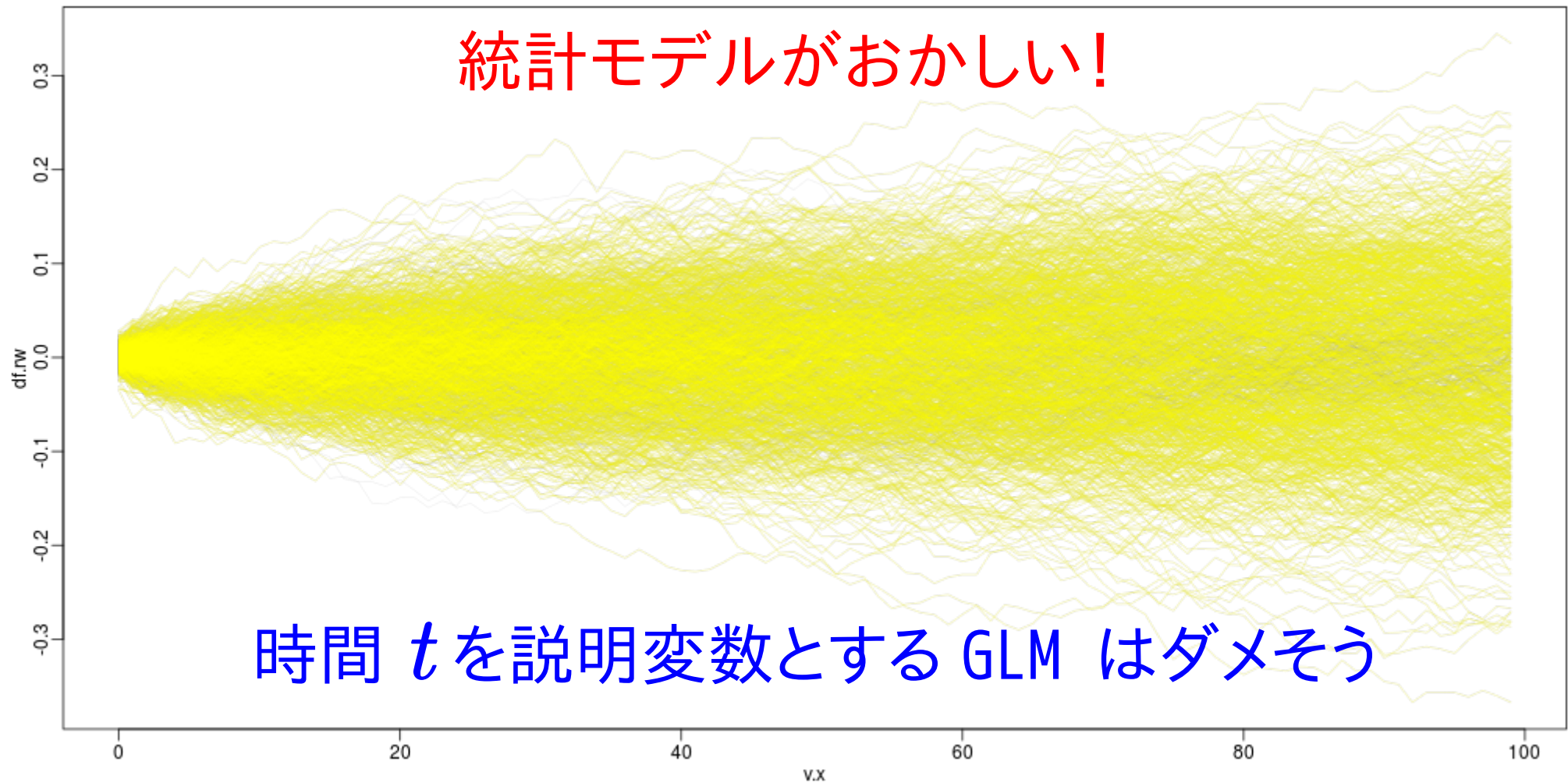
たとえば  $t = 100$  でかなり外れている 50 本





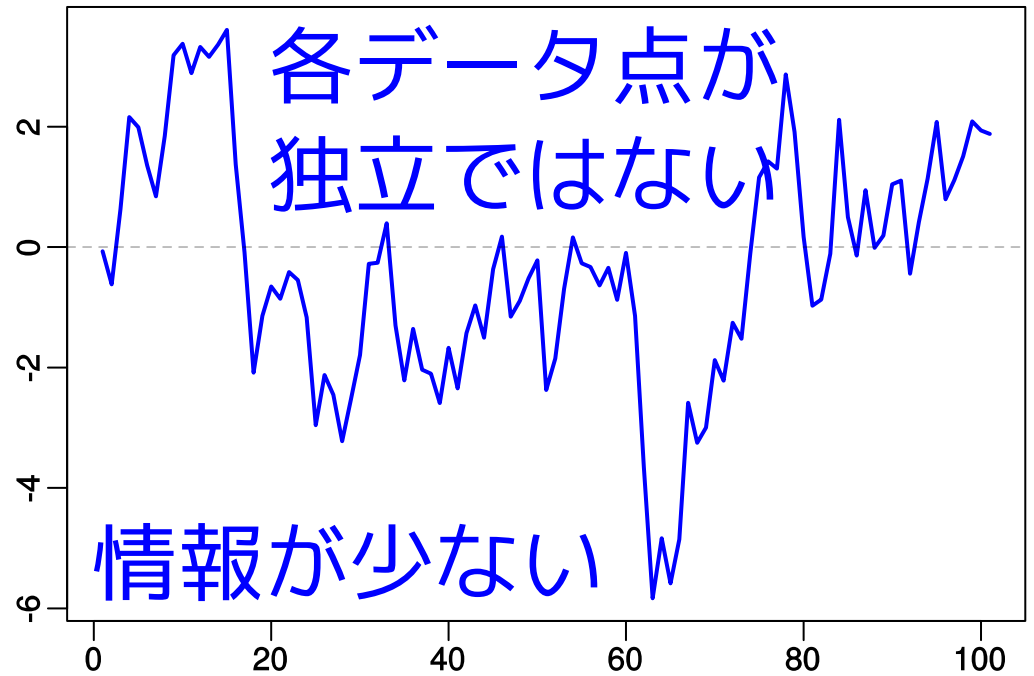
# しかし直線回帰 GLM あてはめると…

ほとんどすべての場合で「ゆーい」！

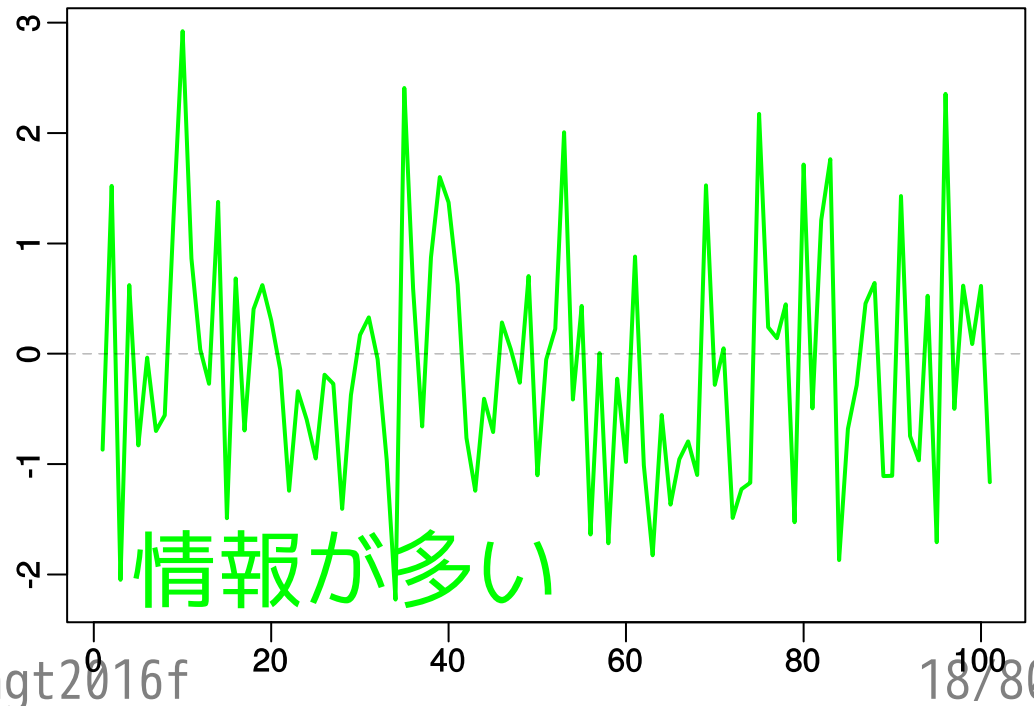


# ちょっとでも傾いてたら「ゆーい」

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?

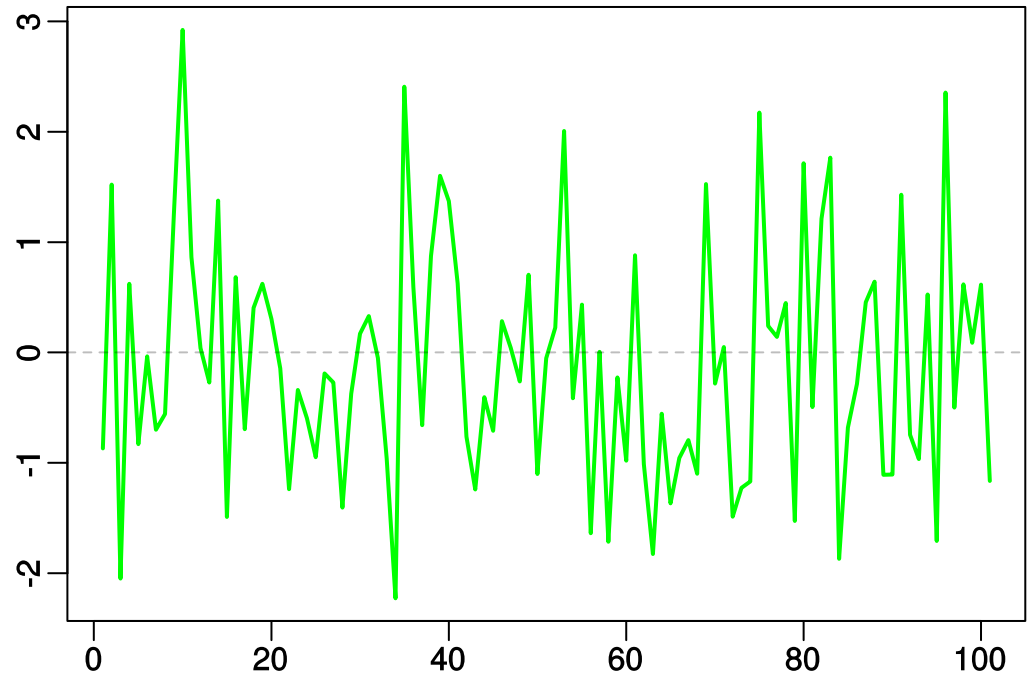
$$\rho_k = \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t) \cdot \text{Var}(y_{t-1})}}$$



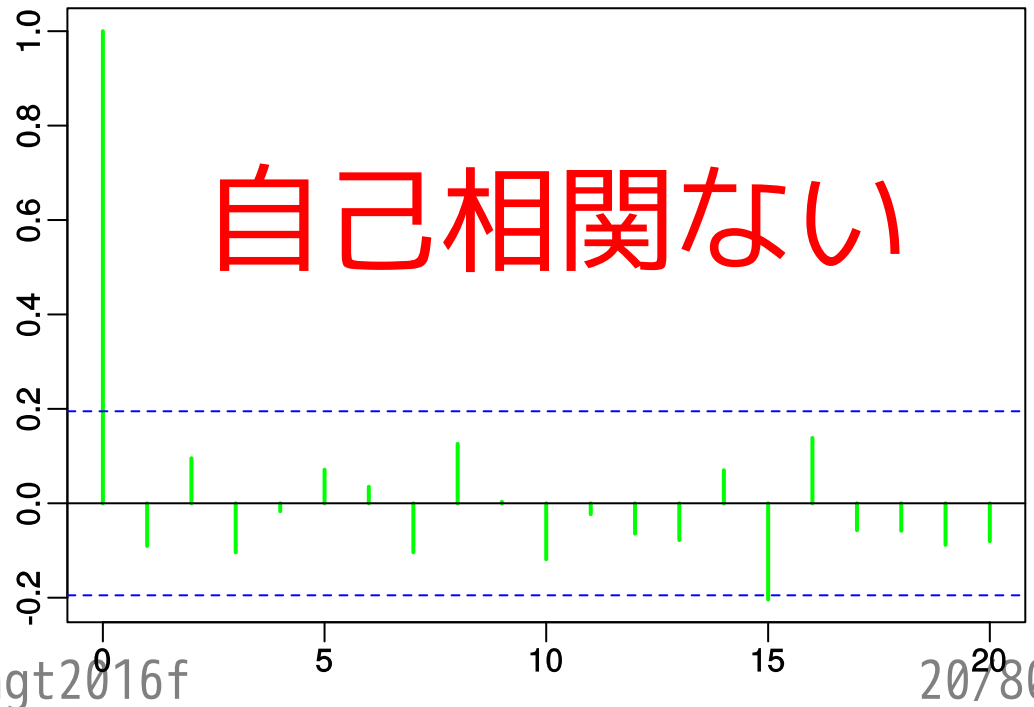
# R の ts クラス: 時系列をあつかう

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

これはたんなる  
100 個の正規乱数

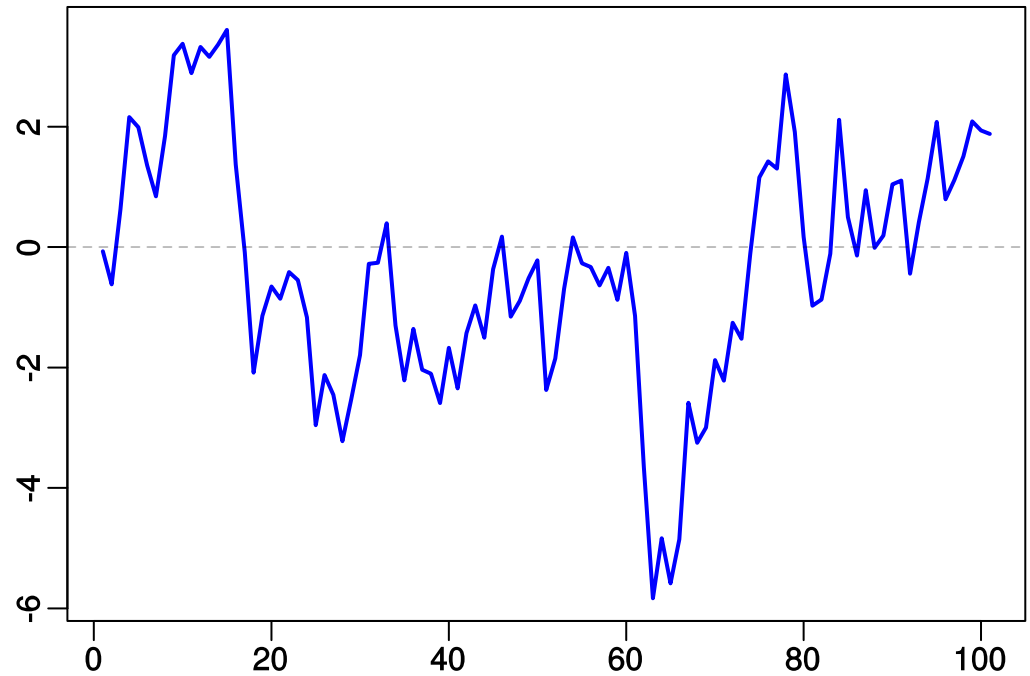


```
plot(acf(ts(Y)))
```

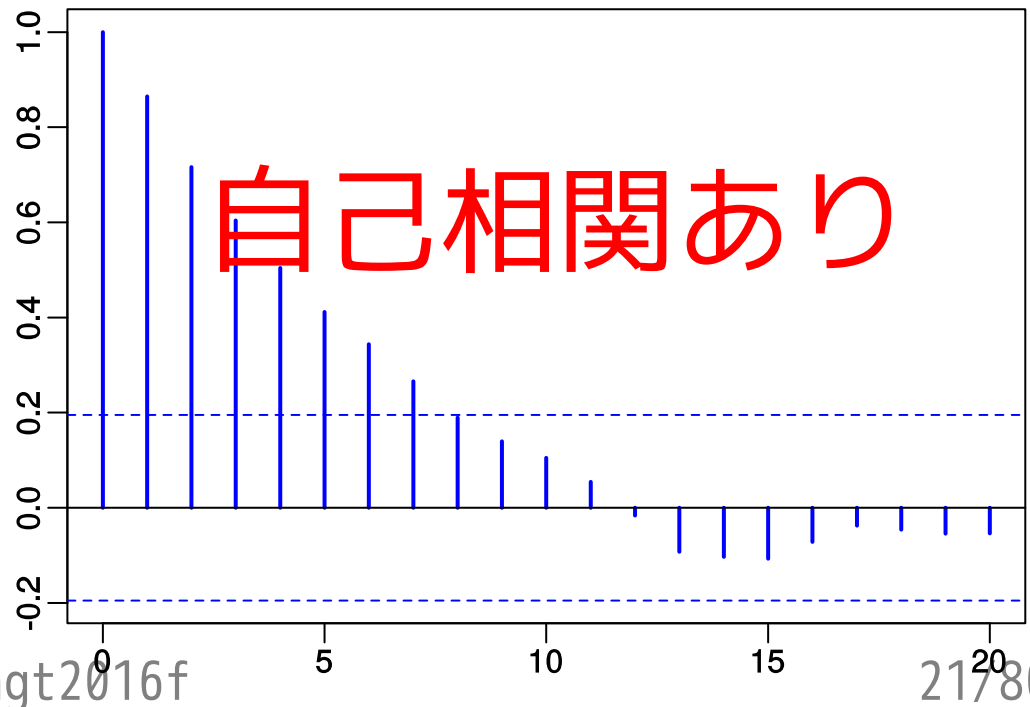


# 自己相関減衰の様子を図示

`plot(ts(Y))`



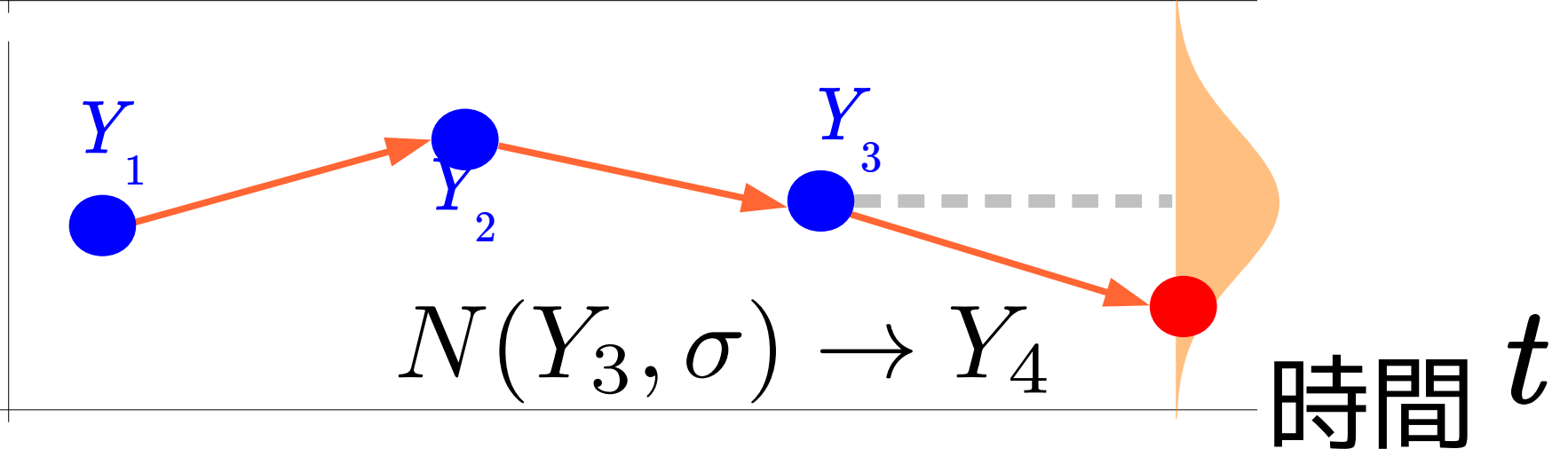
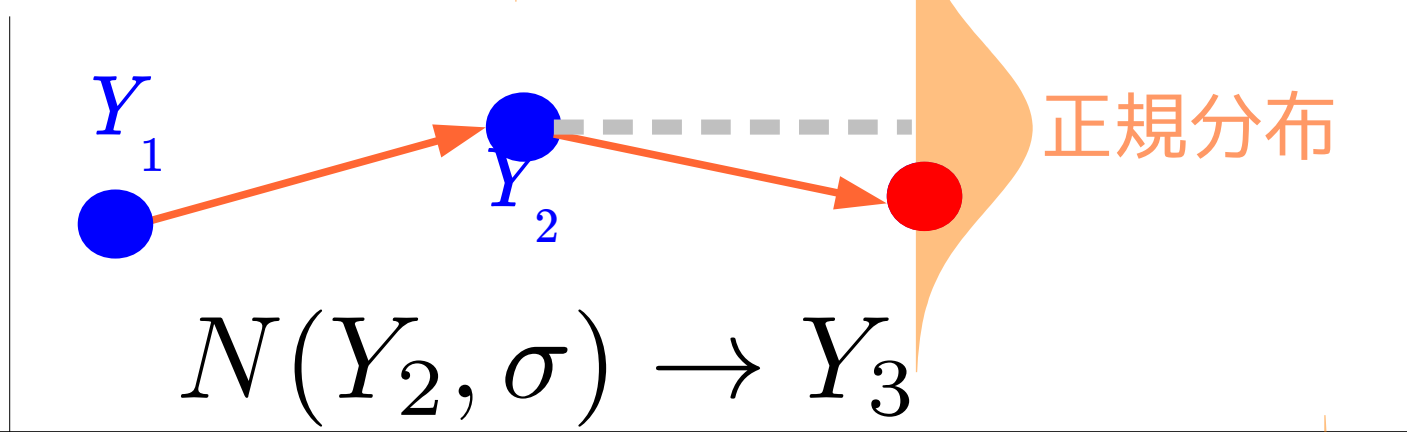
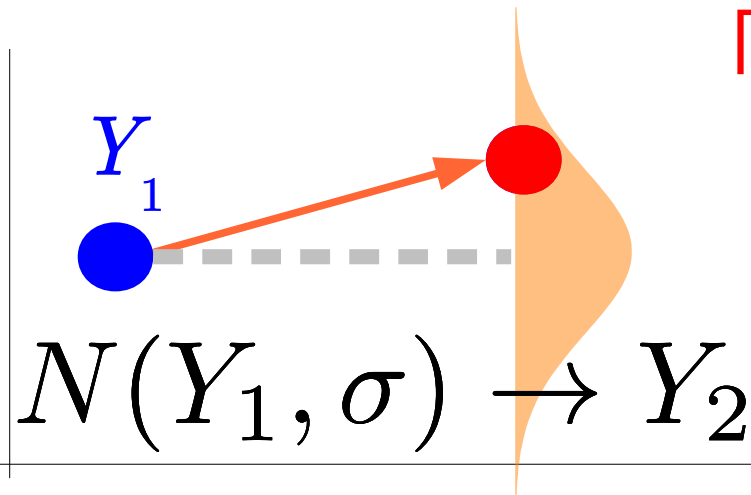
`plot(acf(ts(Y)))`



変数  
 $Y$

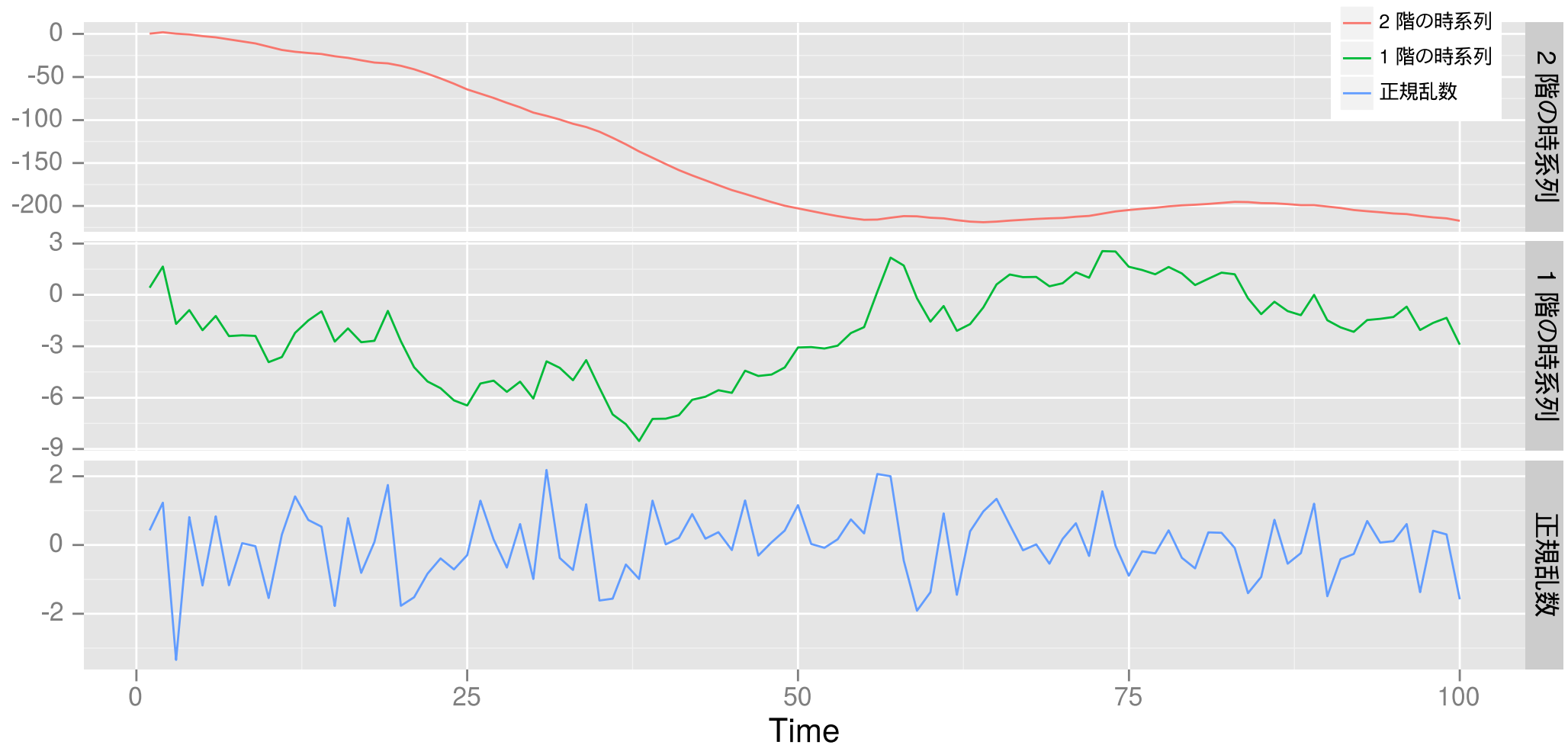
「時間相関がある」とは?

$Y_t$  と  $Y_{t+1}$  は  
似ている!



# 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



## 時間的自己相関

いつも役にたつわけではない?

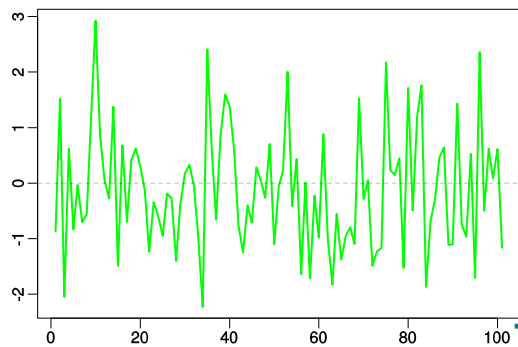
$$\rho_k = \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t) \cdot \text{Var}(y_{t-1})}}$$



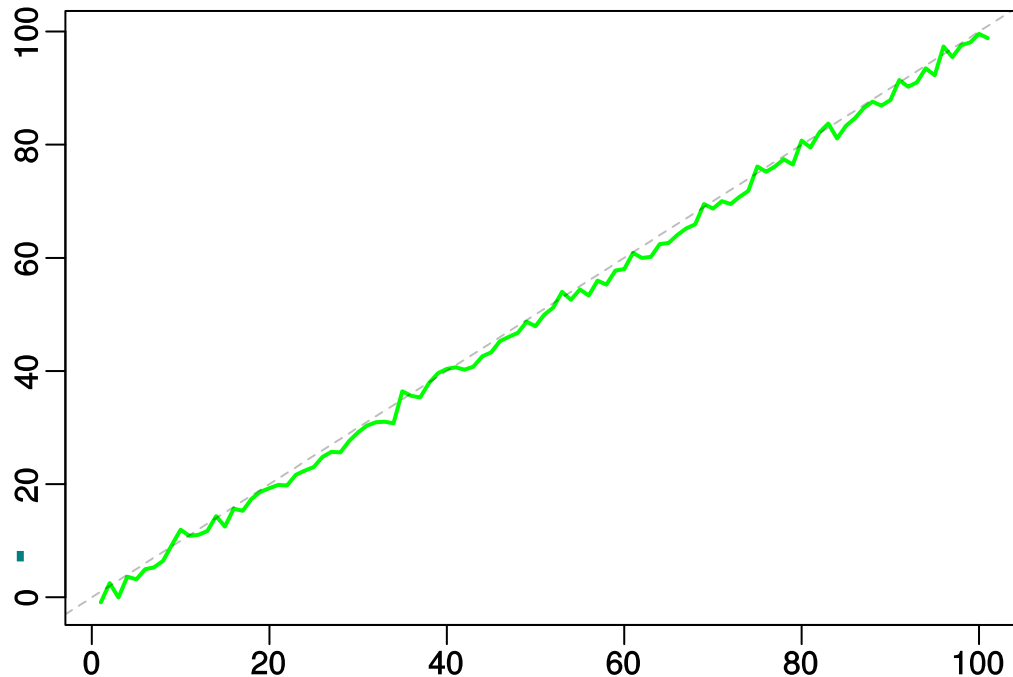


# 各点独立のデータをナナメにすると？

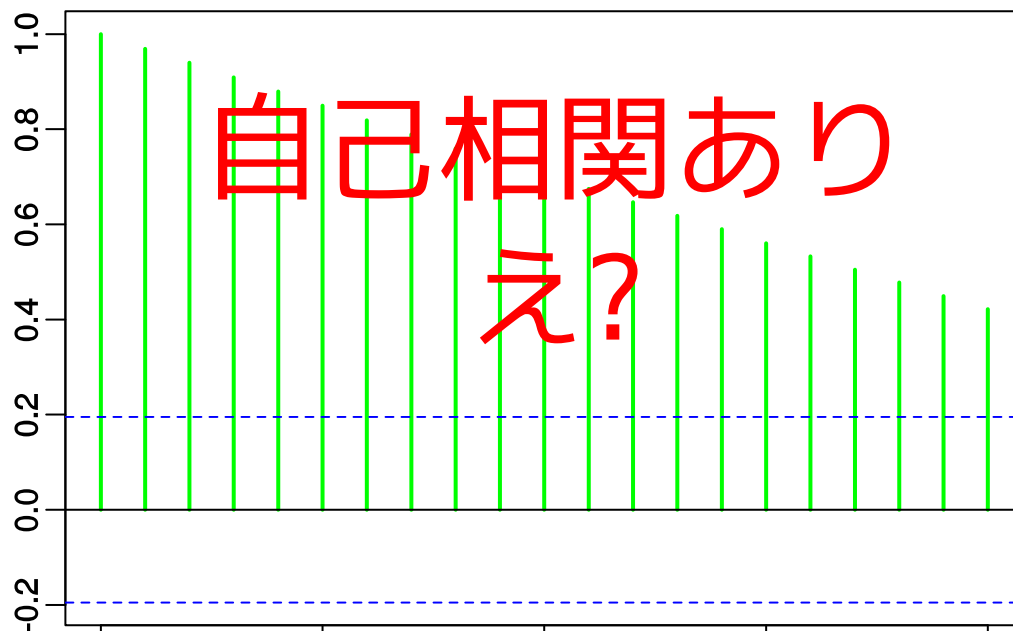
`plot(ts(Y))`



これを  
ナナメに  
したもの  
なんだけど...



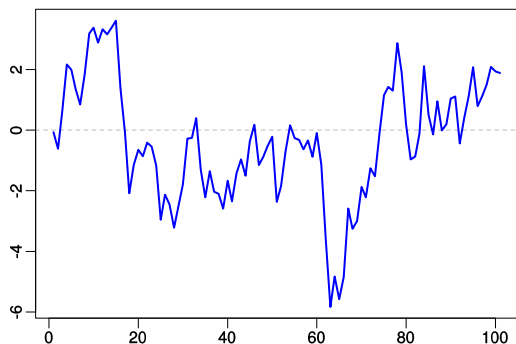
`plot(acf(ts(Y)))`



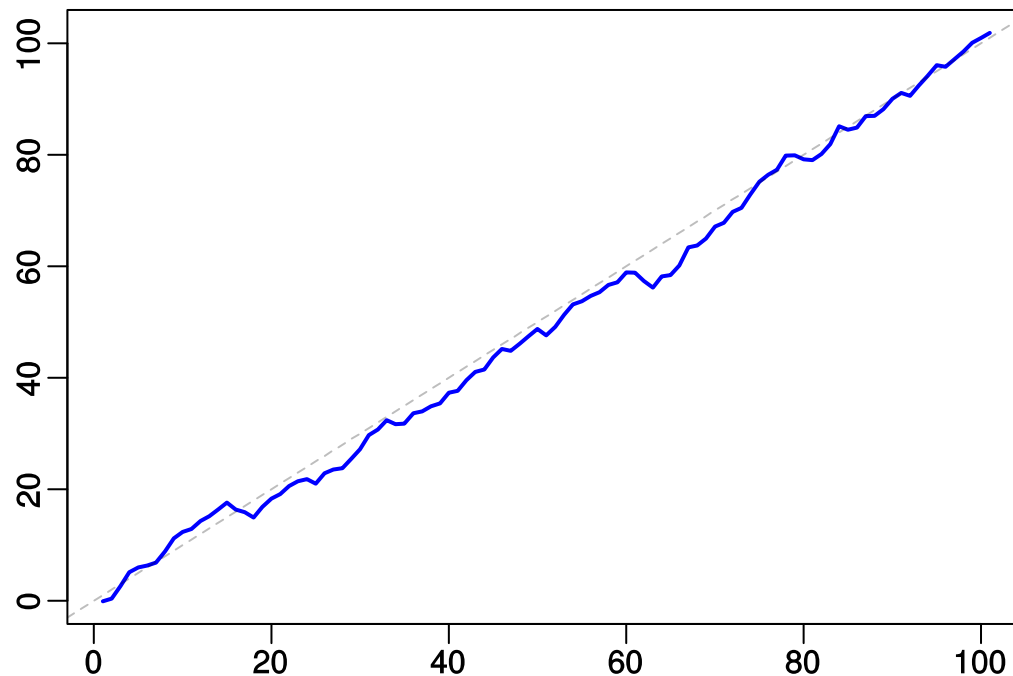
自己相関あり  
え？

# 各点独立のデータをナナメにすると?

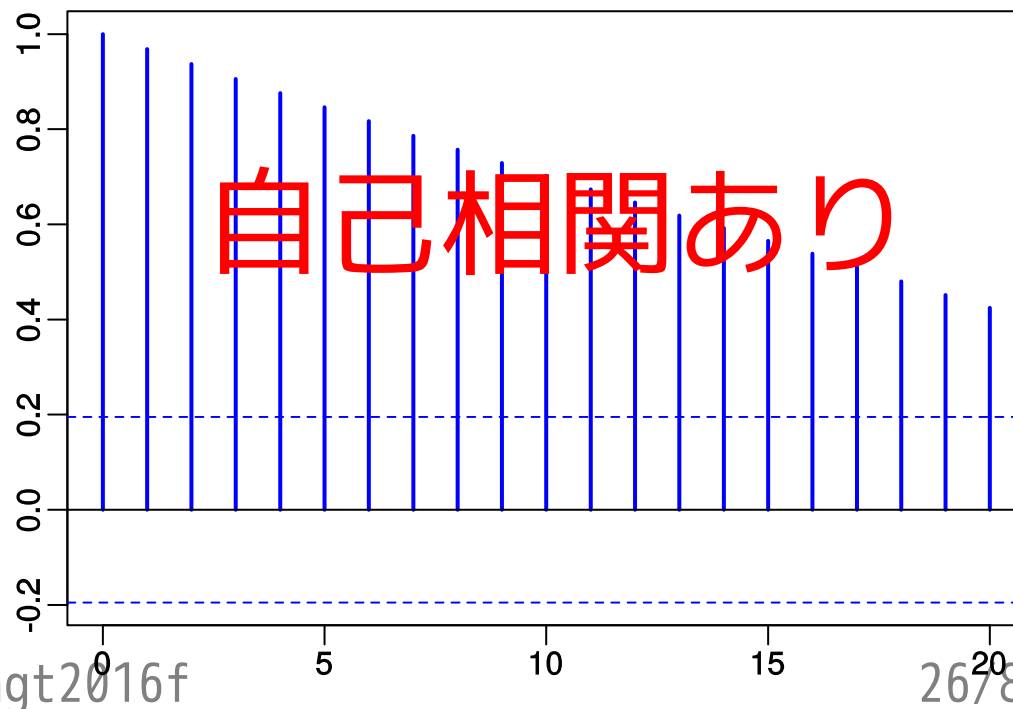
`plot(ts(Y))`



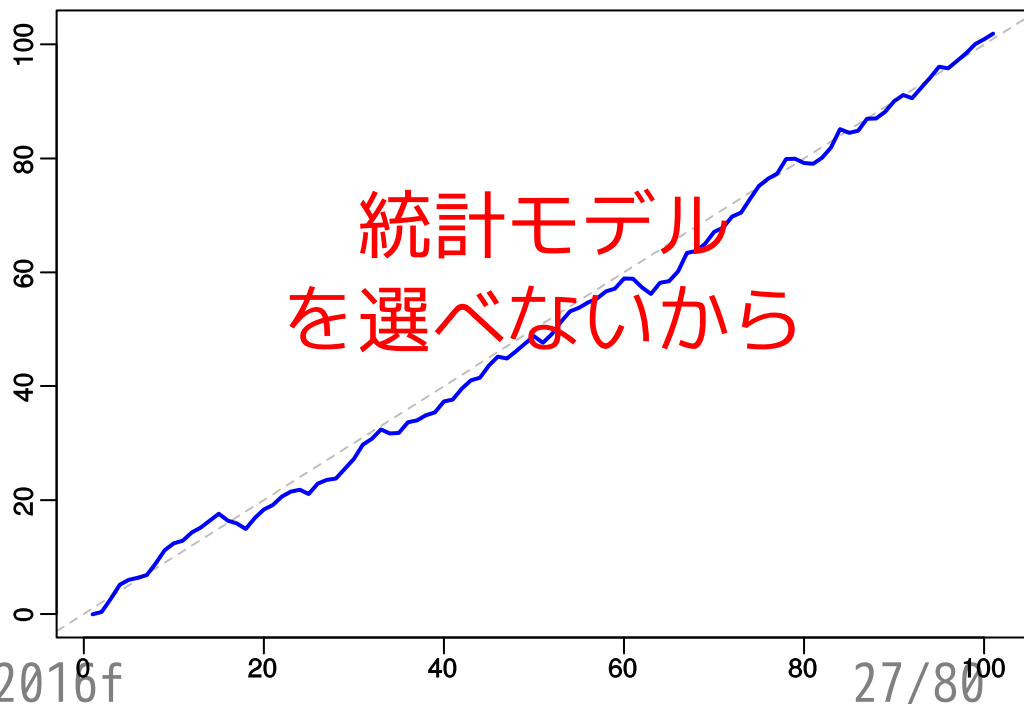
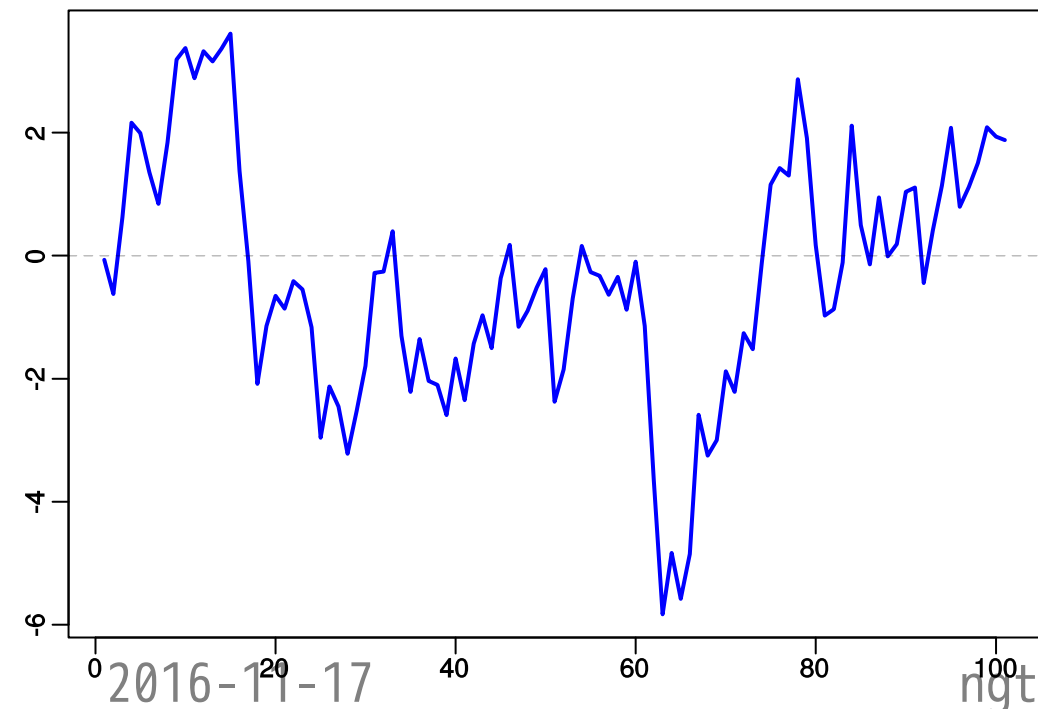
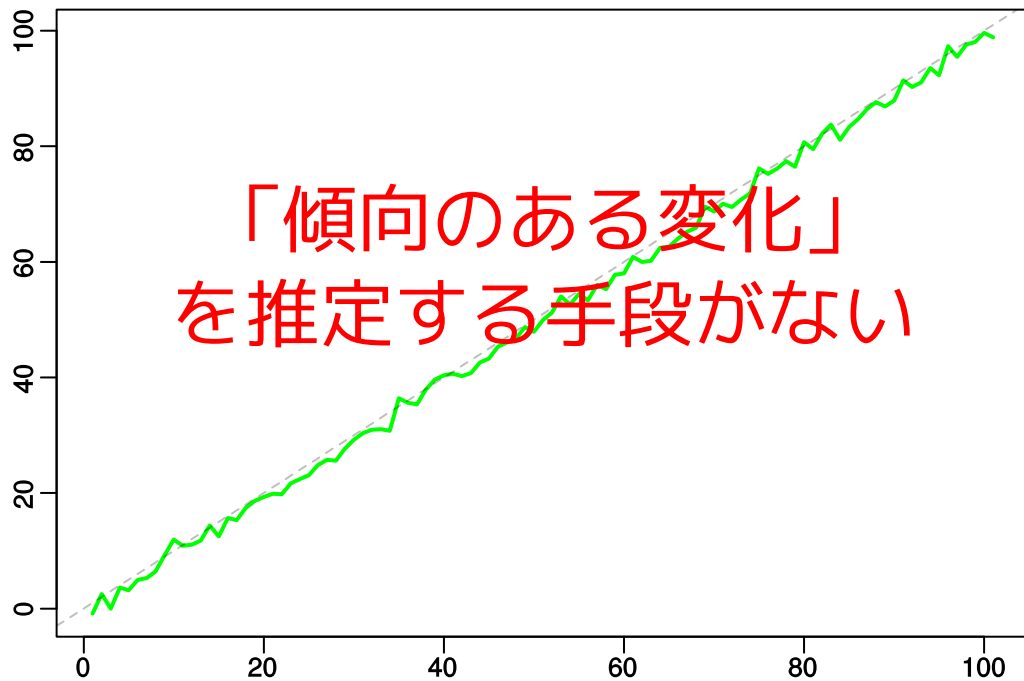
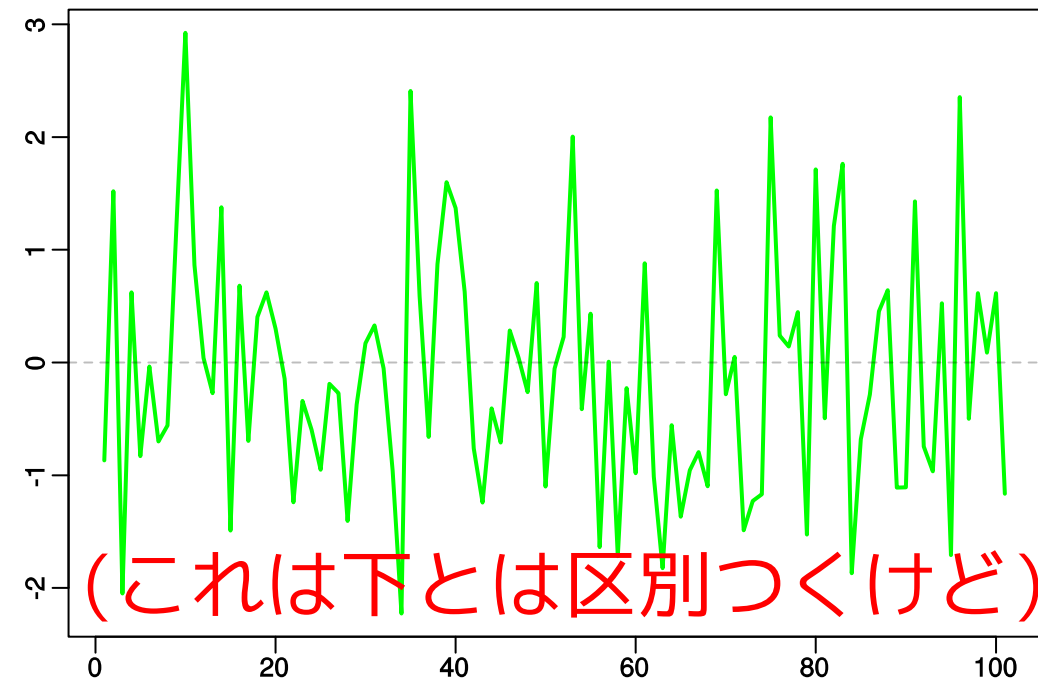
これを  
ナナメに  
したもの



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

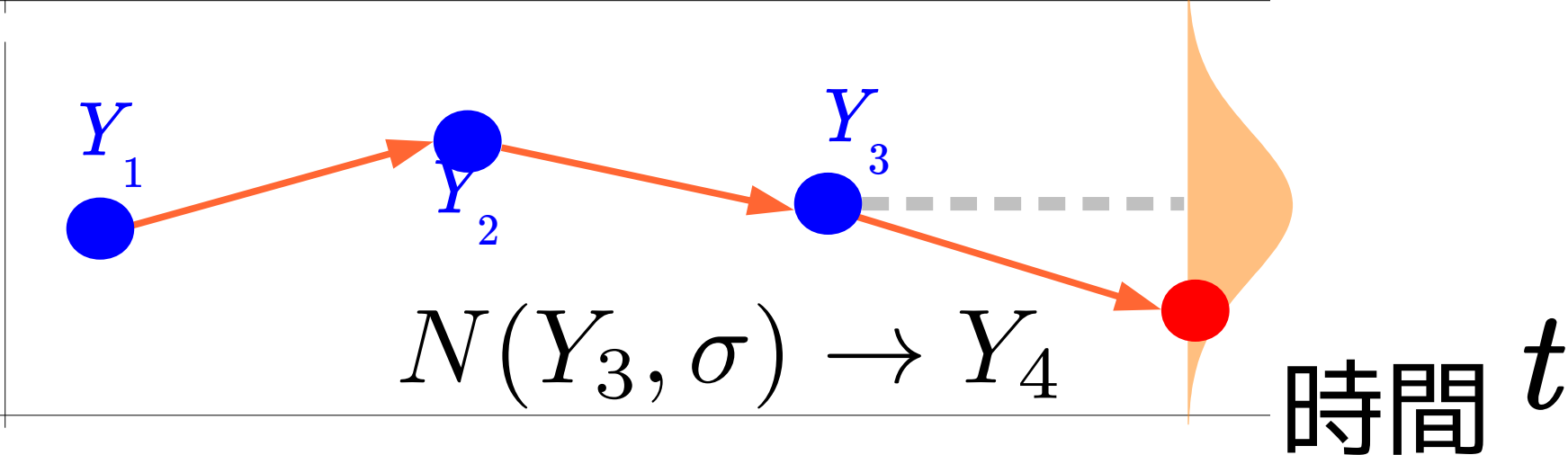
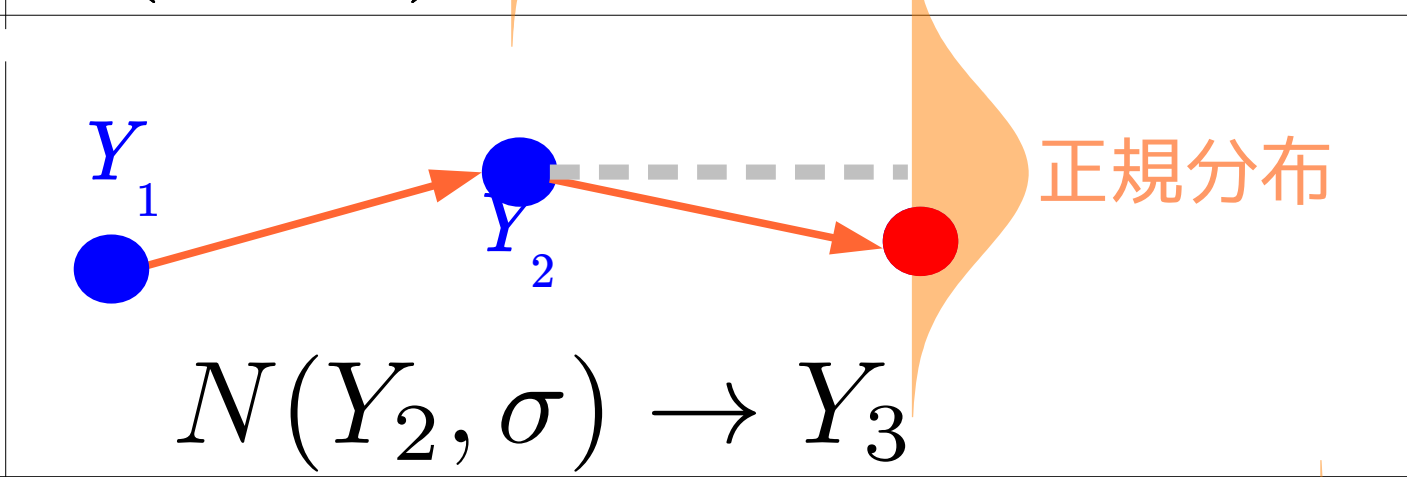
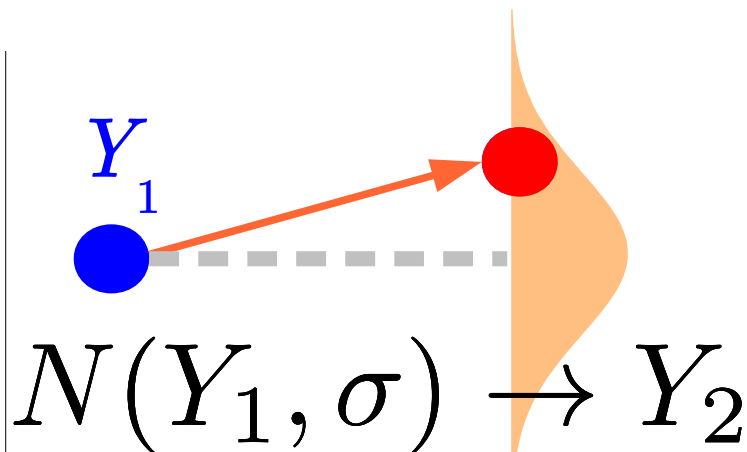


# 自己相関係数みても区別がつかない



変数  
 $Y$

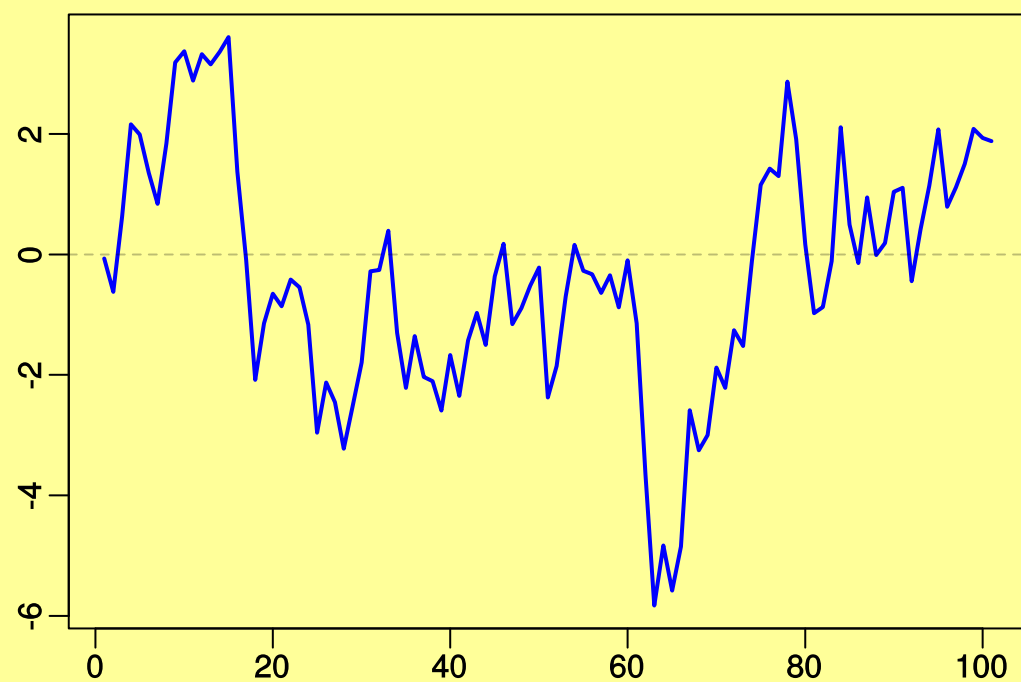
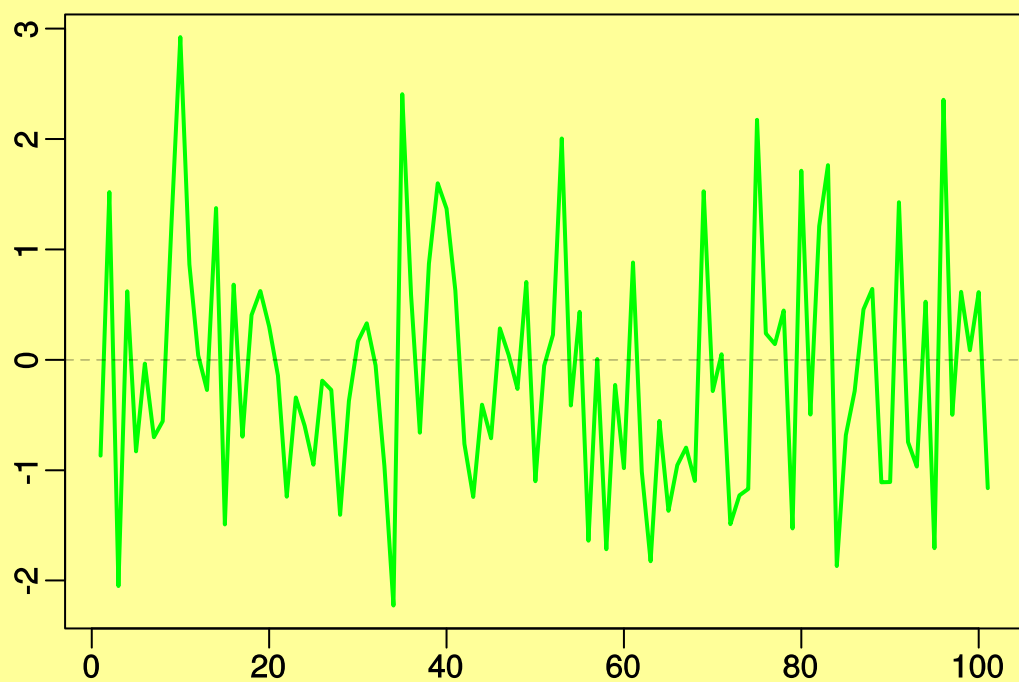
ランダムウォーク  
もっとも単純な  
モデル



# 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？



時系列データ解析の教科書，ねえ……

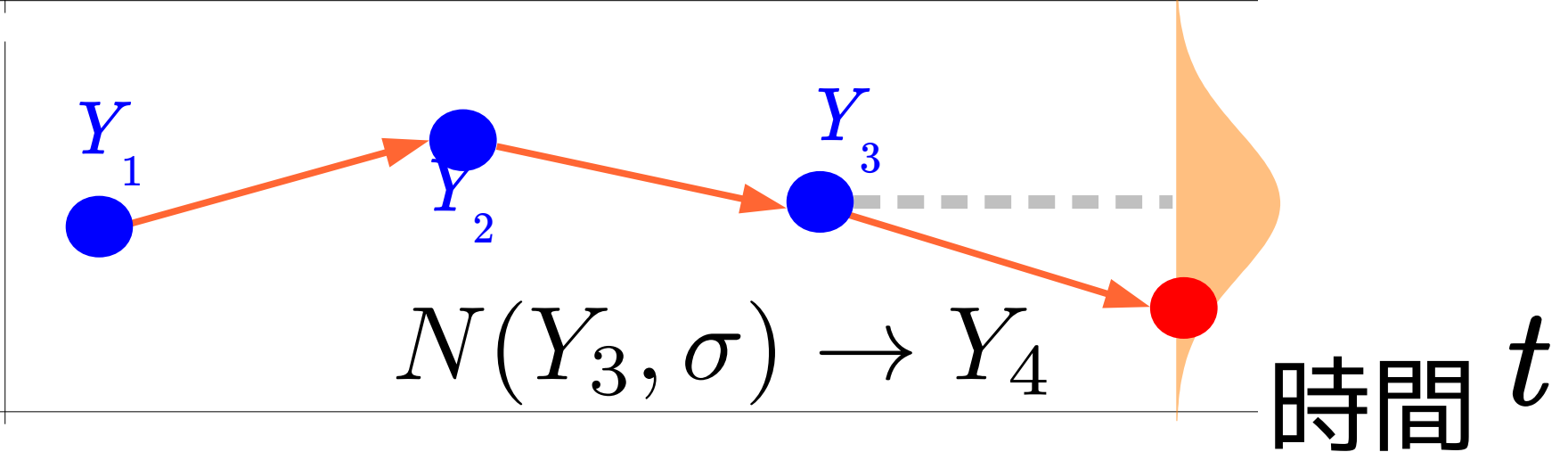
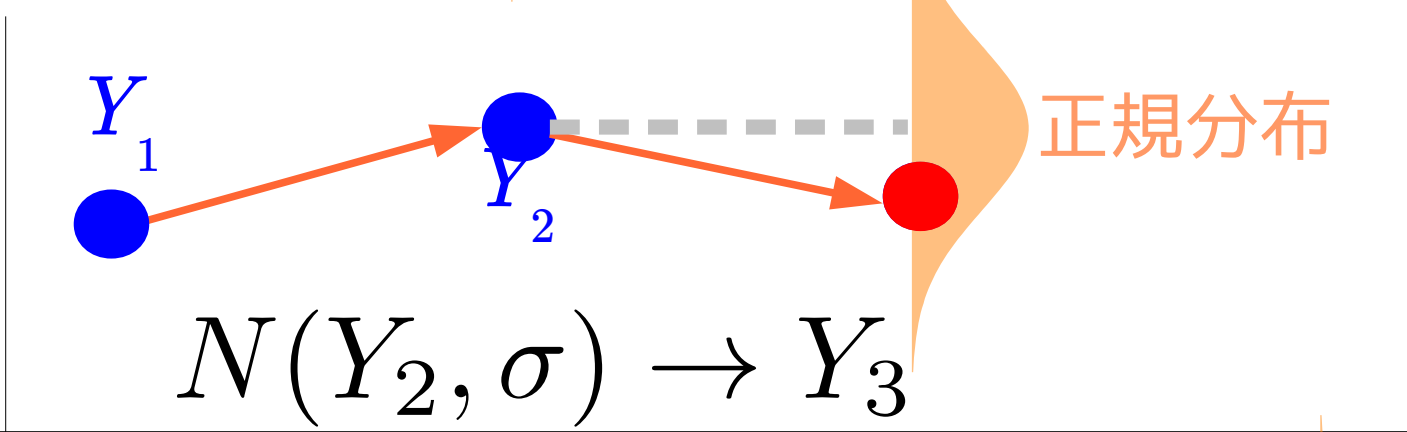
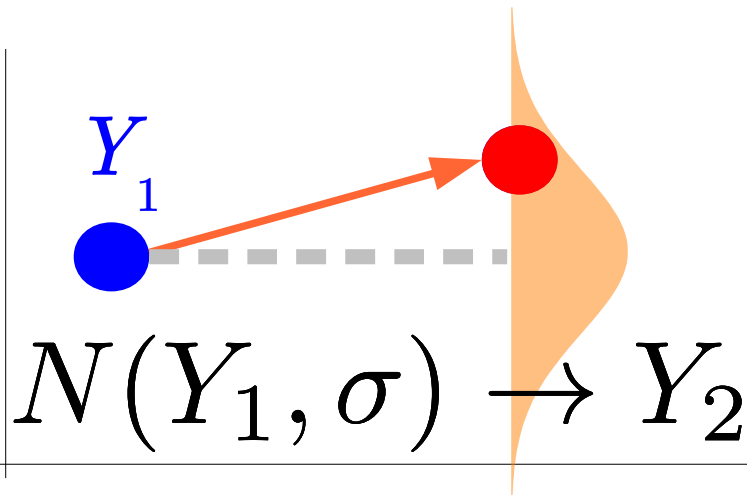
- モデルがあれこれ多すぎる
- 経済学よりのモデルばかり
- なんでも正規分布

なんとかならないかな？

**状態空間モデル**，どうでしょう？

変数  
 $Y$

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 $\sigma$ をもつ

観測の誤差

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

観測データ  $Y_1$

$Y_2$

$Y_3$

$y_1$

$y_2$

$y_3$

$y_4$

$$N(y_t, \sigma_1) \rightarrow y_{t+1}$$

状態変数の変化

時間  $t$

観測できない世界 (状態空間)

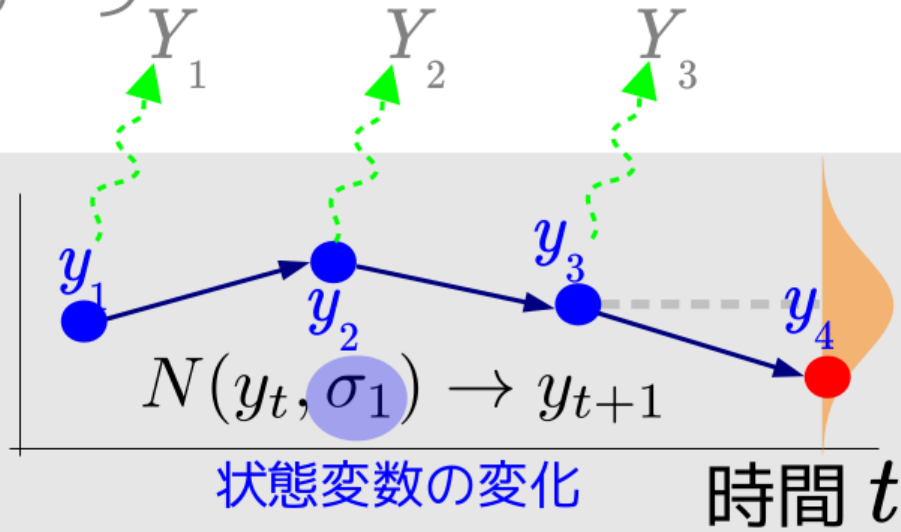


# 状態空間モデル

観測の誤差

$$N(y_t, \sigma_2) \rightarrow Y_t \quad \text{二種類の } \sigma \text{ をもつ}$$

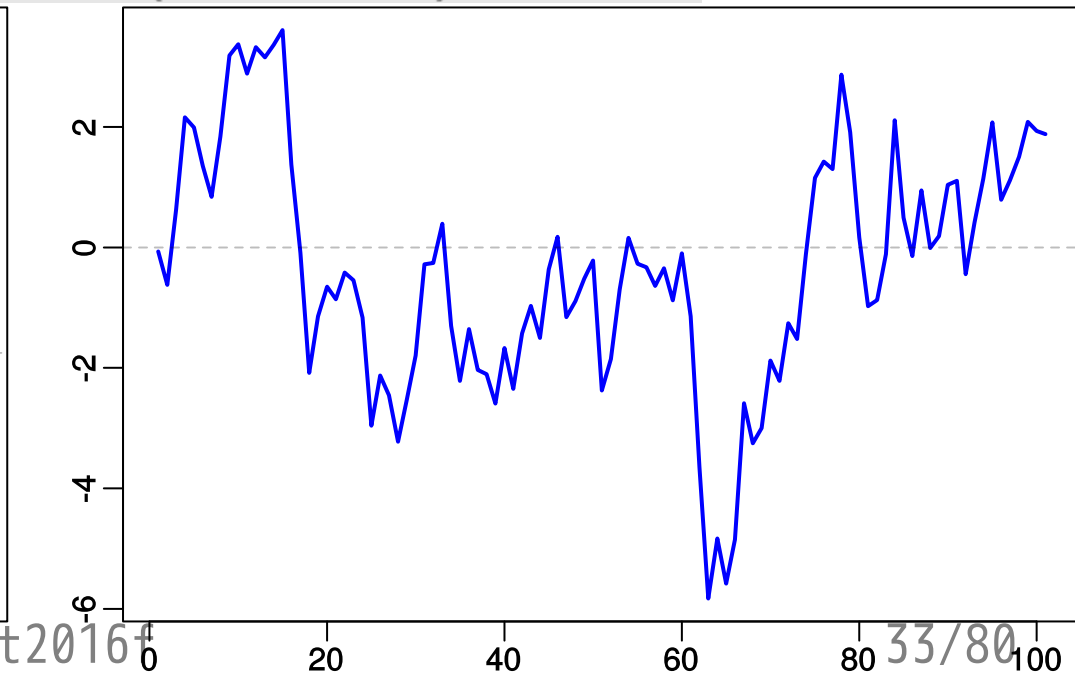
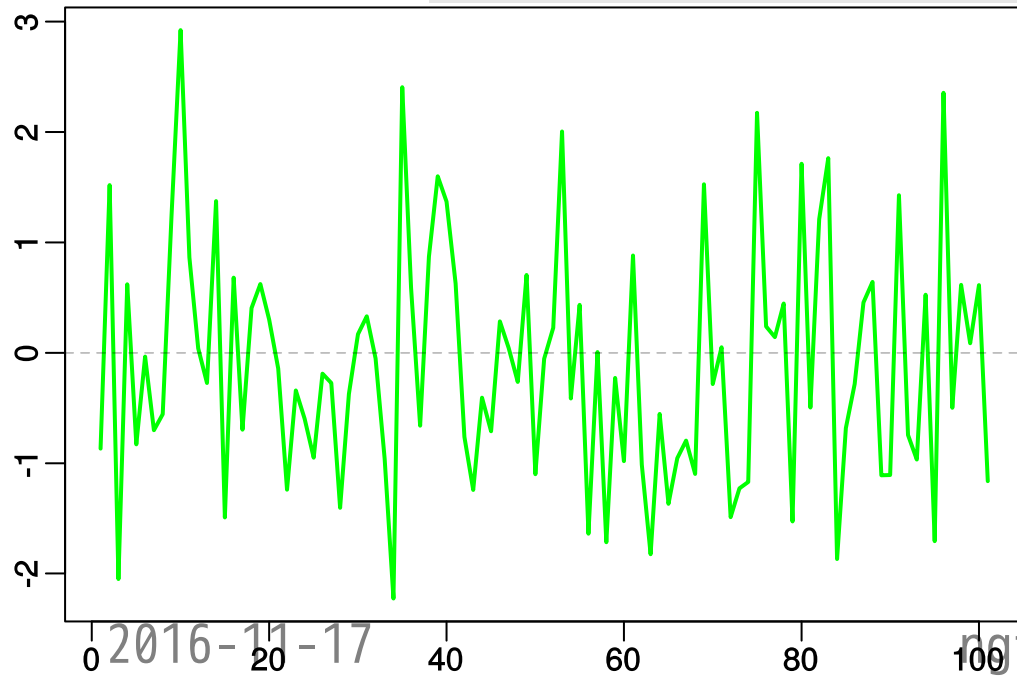
観測データ



$\sigma_2$  大  
 $\sigma_1$  小

$\sigma_2$  小  
 $\sigma_1$  大

観測できない世界 (状態空間)

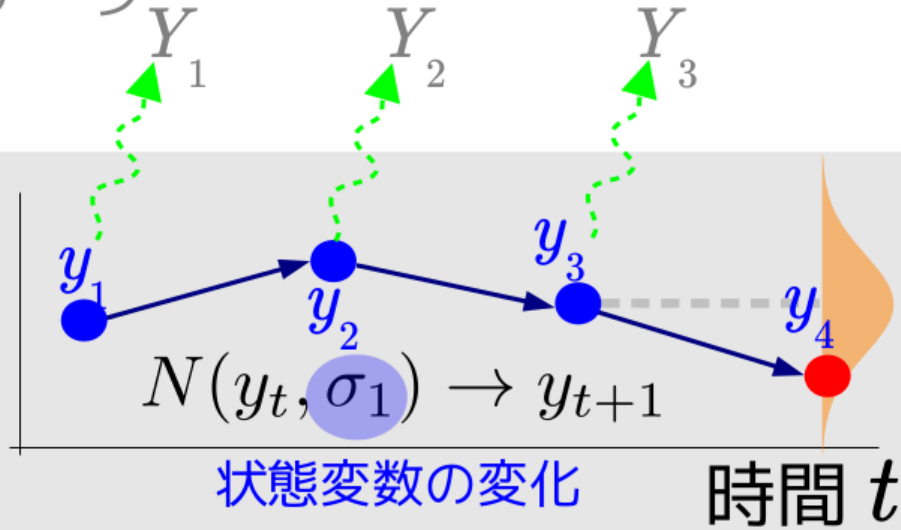


# 状態空間モデル

観測の誤差

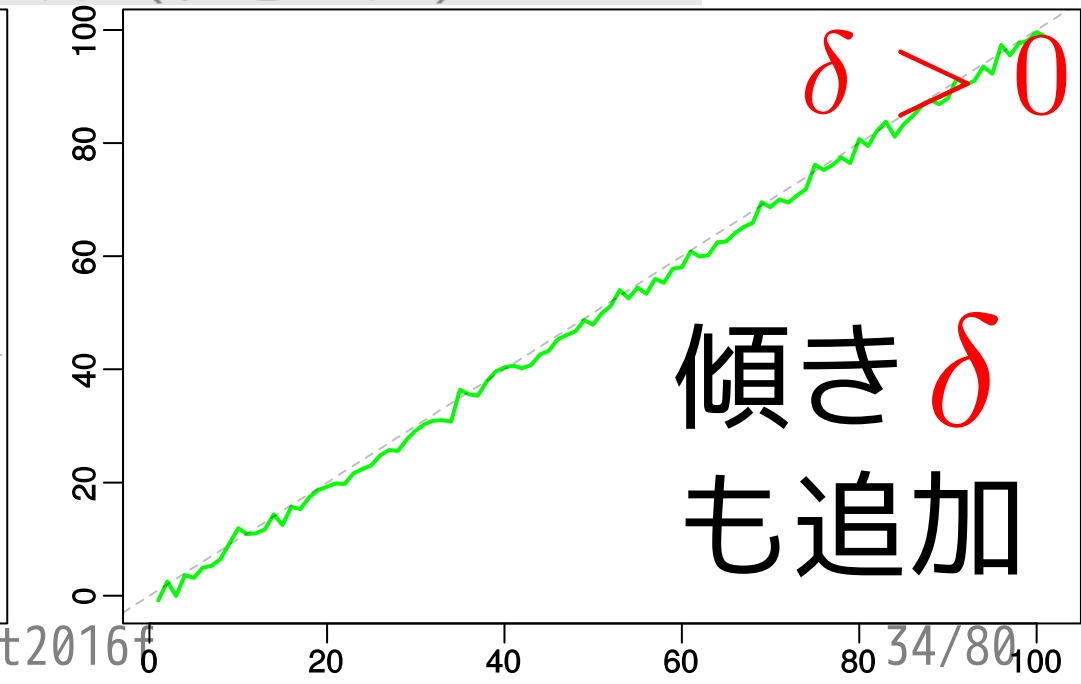
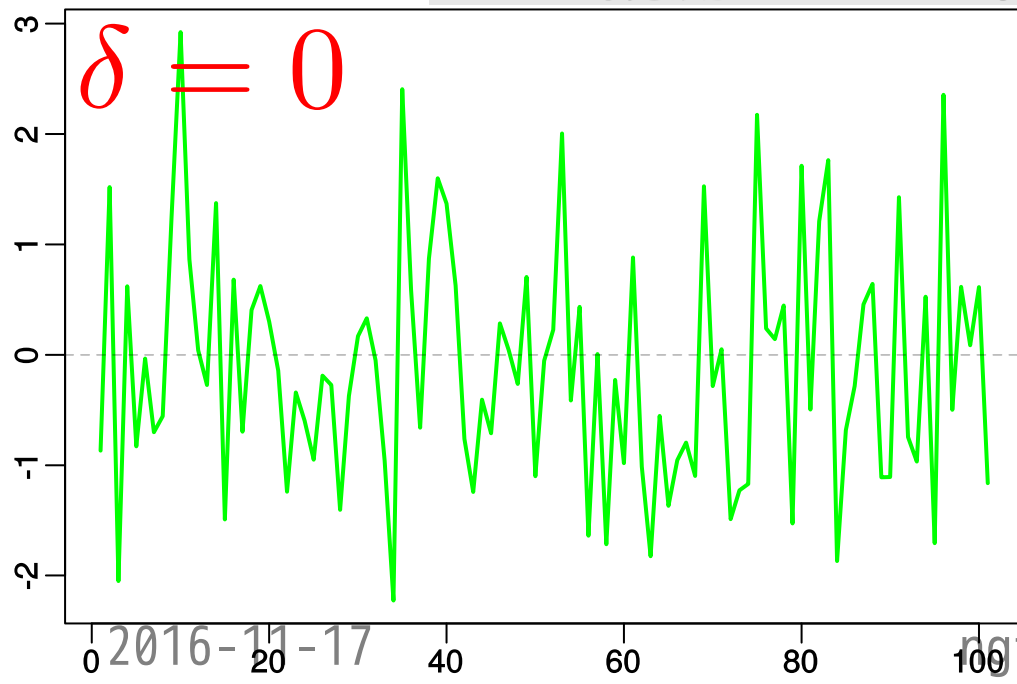
$$N(y_t, \sigma_2) \rightarrow Y_t \quad \text{二種類の } \sigma \text{ をもつ}$$

観測データ



観測できない世界 (状態空間)

$\sigma_2$  大  
 $\sigma_1$  小

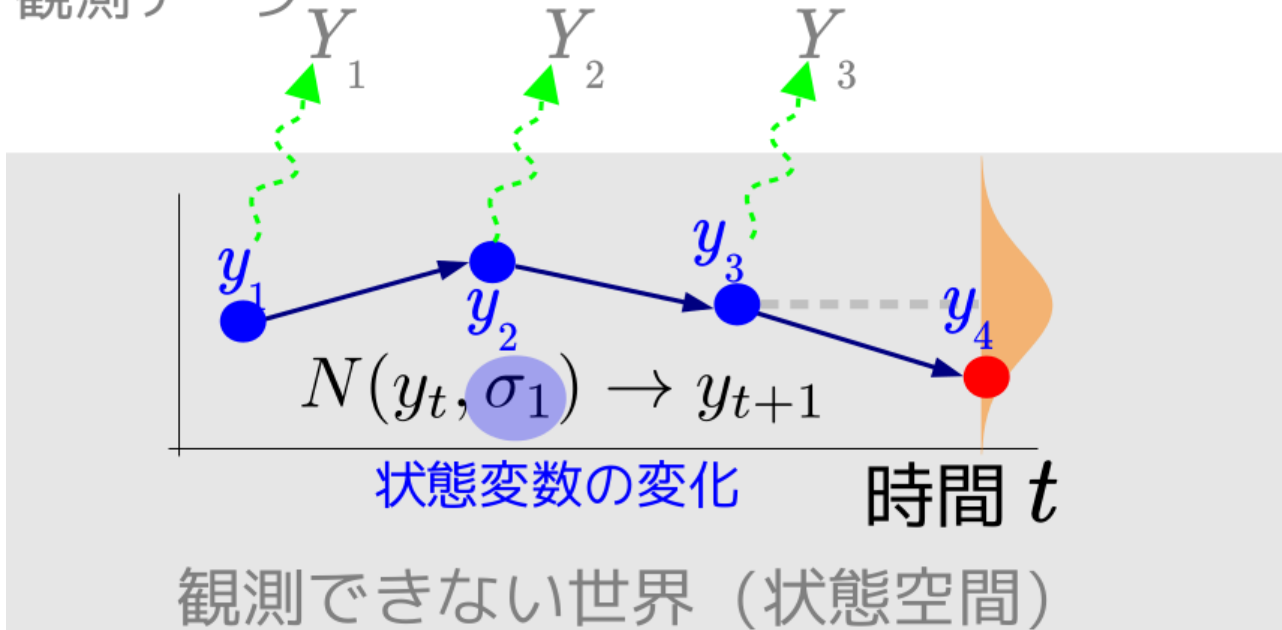


# 状態空間モデル

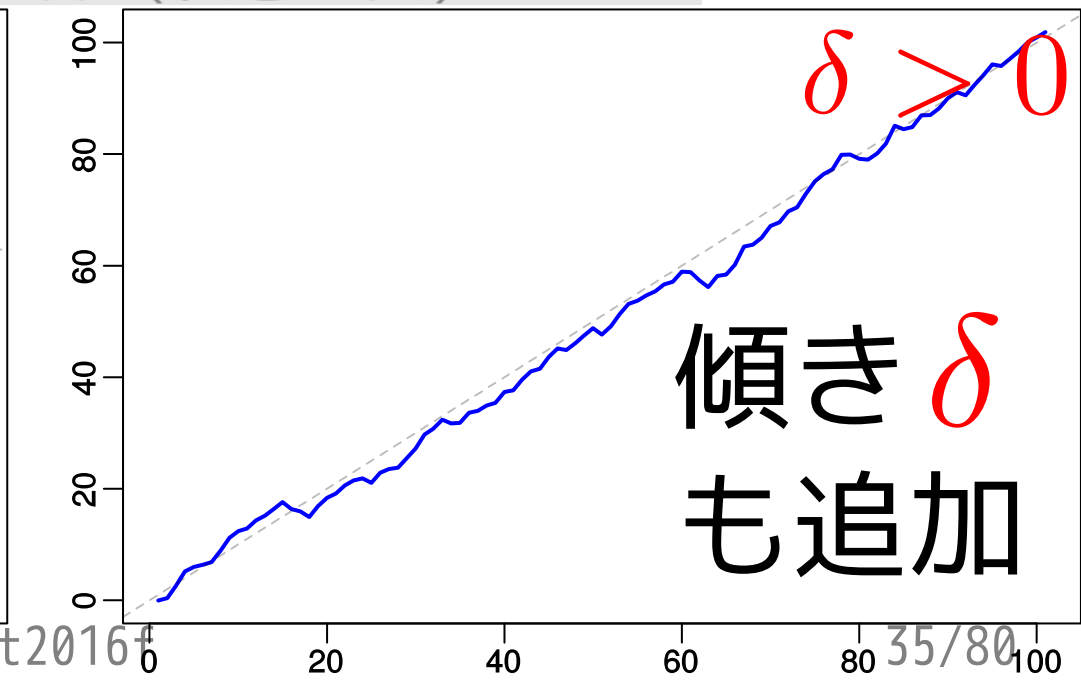
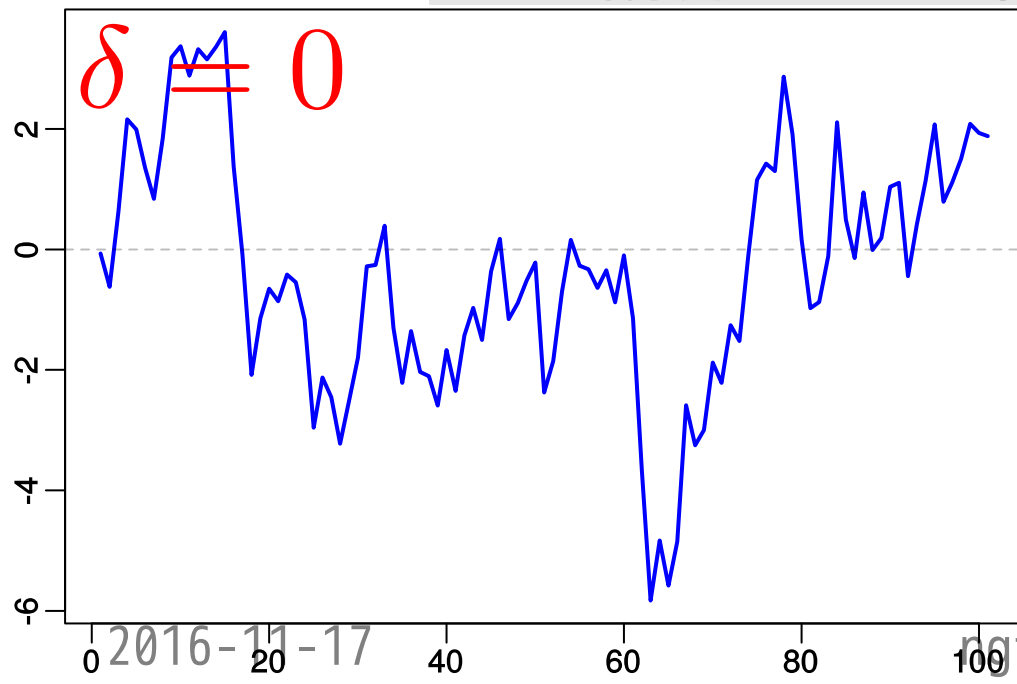
観測の誤差

$$N(y_t, \sigma_2) \rightarrow Y_t \quad \text{二種類の } \sigma \text{ をもつ}$$

観測データ



$\sigma_2$  小  
 $\sigma_1$  大

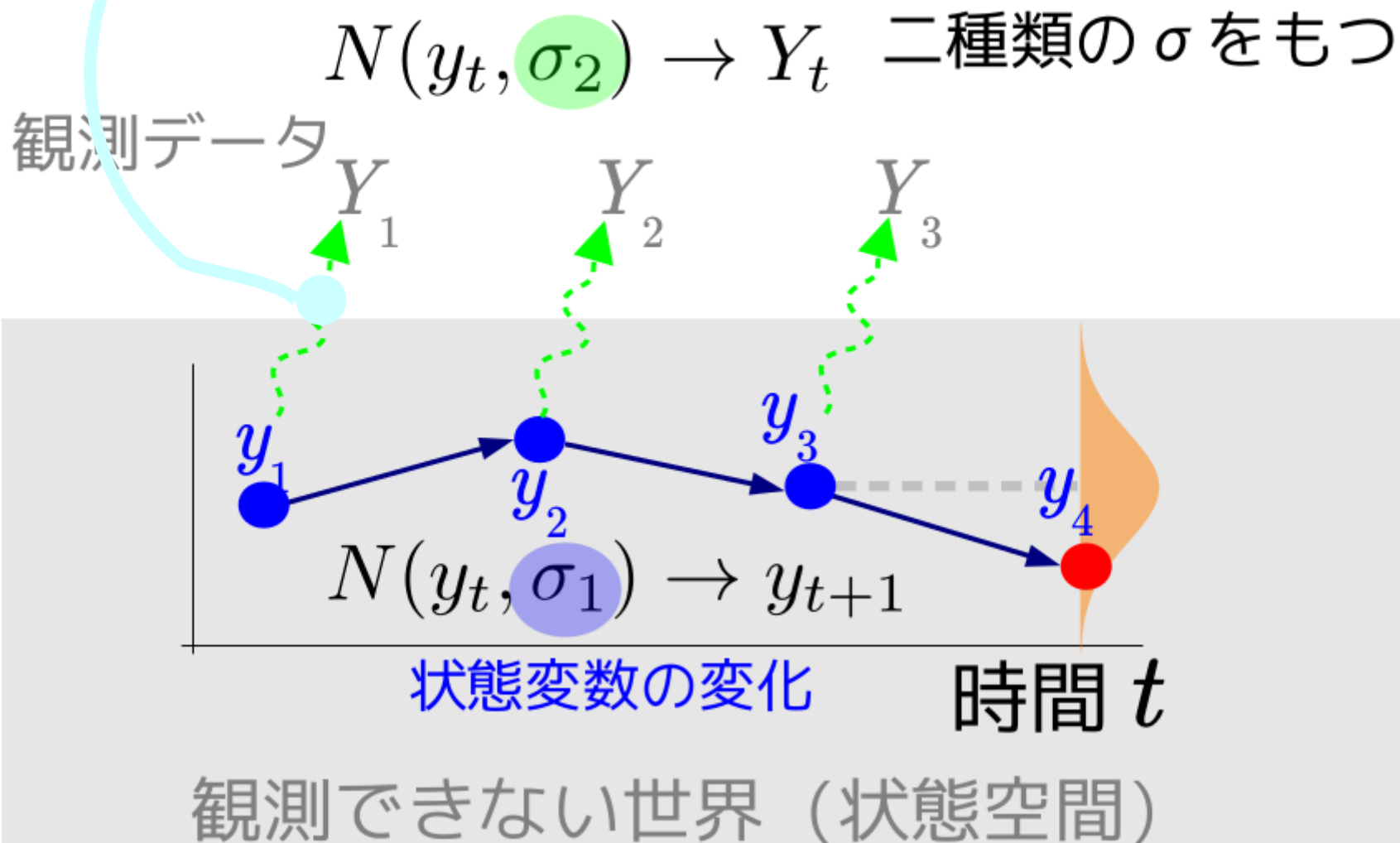


# 状態空間モデル + GLM

この部分にポアソン分布や  
二項分布をいれる

誤差

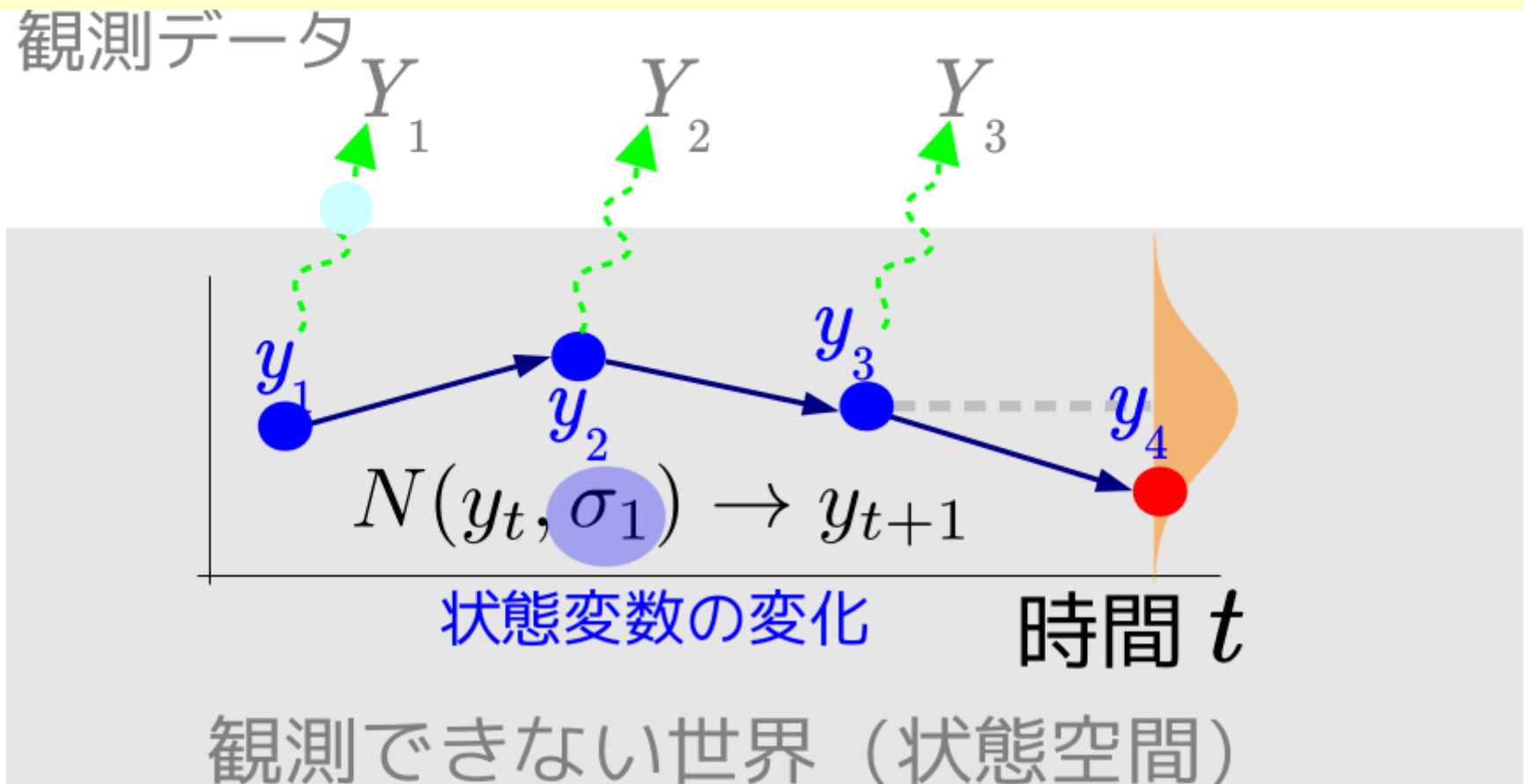
状態空間モデル



# 状態空間モデル + GLM

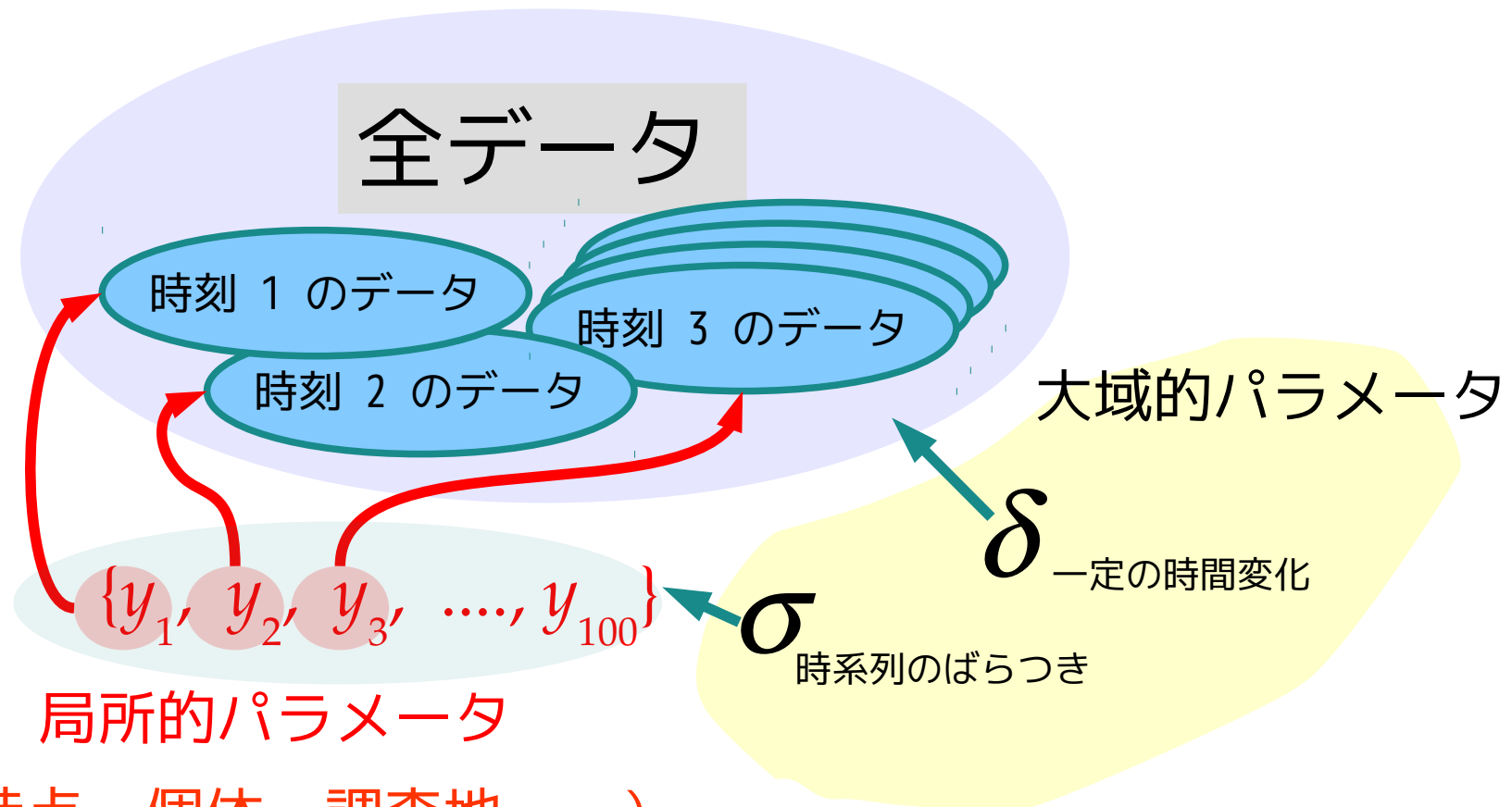
他にも季節変動などを入れることができます

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……)

# どうやってモデルをあてはめる？



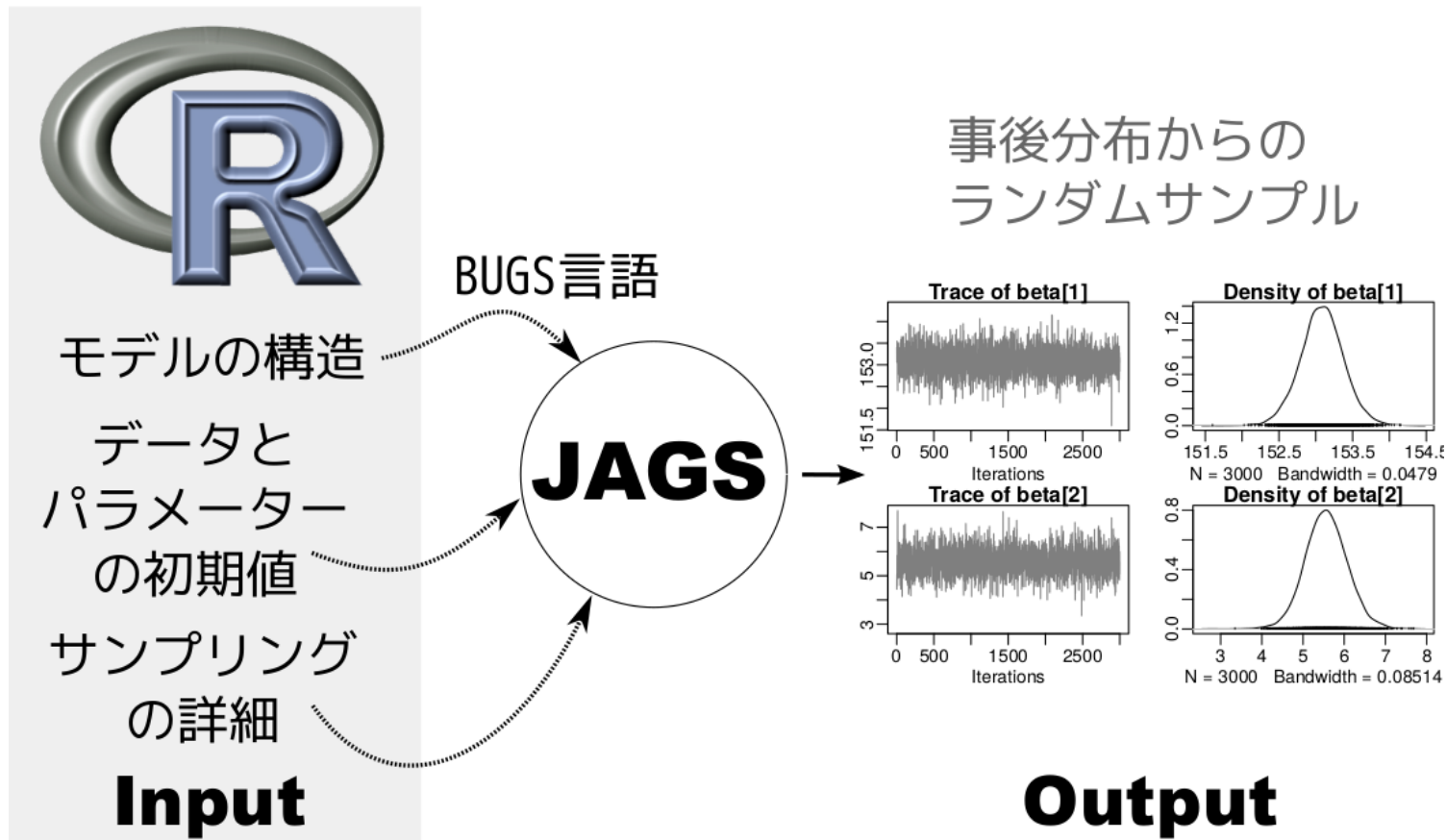
R の状態空間モデルの  
package いろいろある

`library(dlm)`

`library(KFAS)`

(けっこうめんどう?)

# 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) {
```

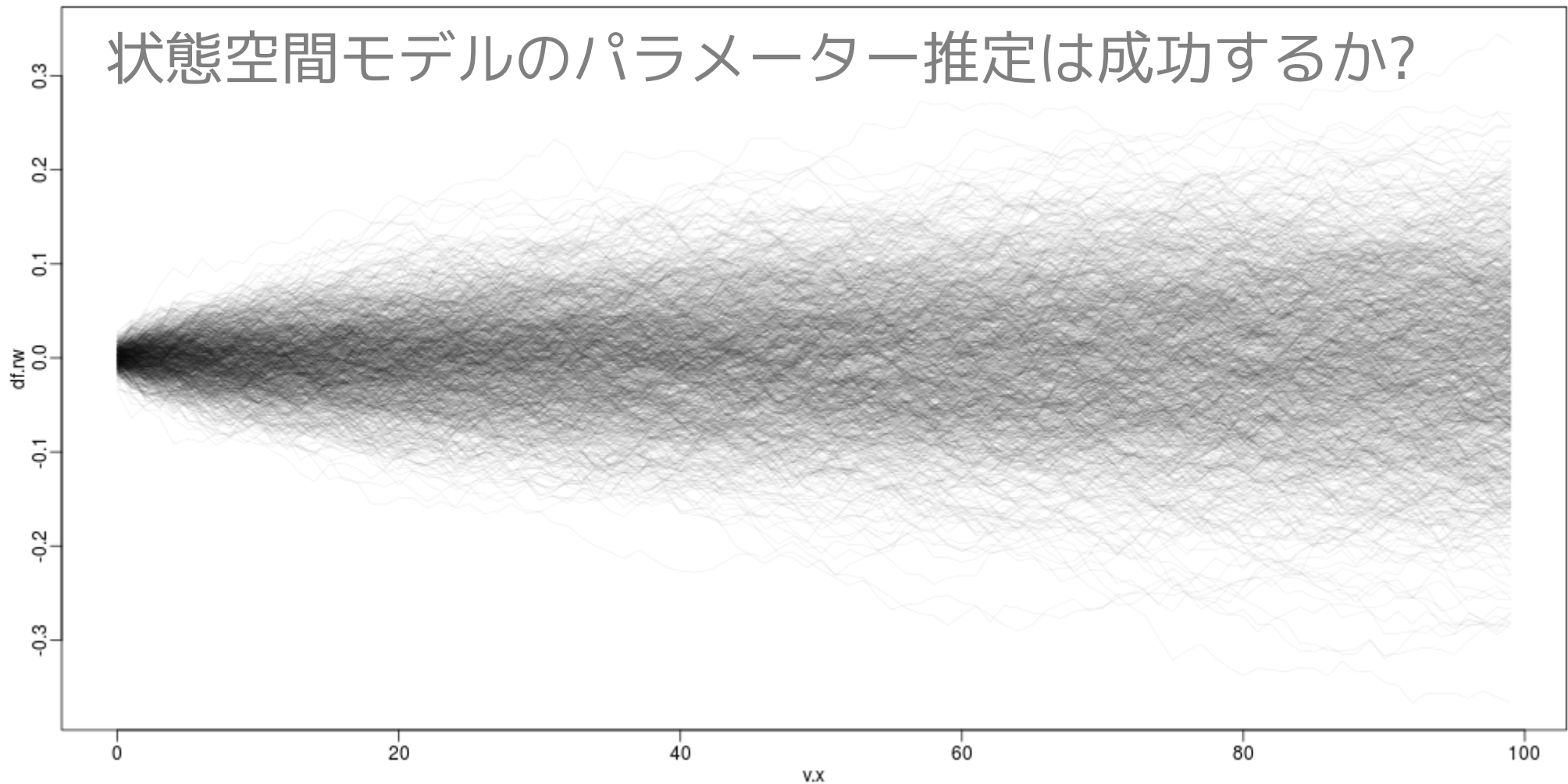
```
    tau[k] <- 1 / (s[k] * s[k])
```

```
    s[k] ~ dunif(0, 10000)
```

```
  }
```

# 1000 個の架空データを推定

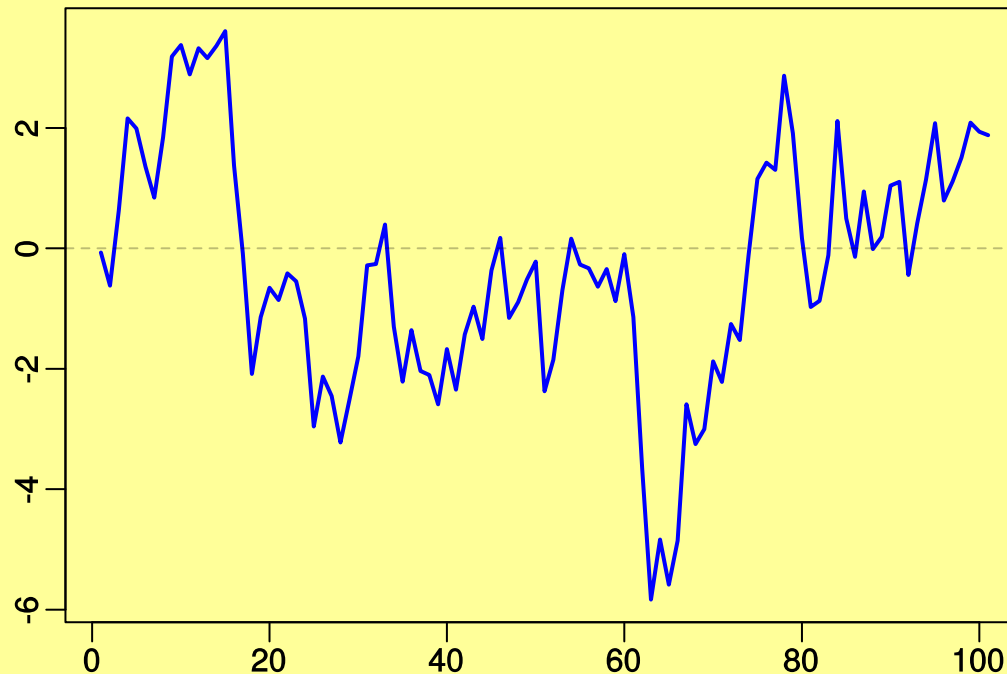
いろいろなランダムウォークが生成される



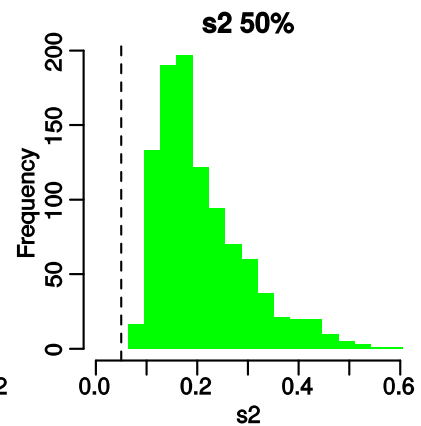
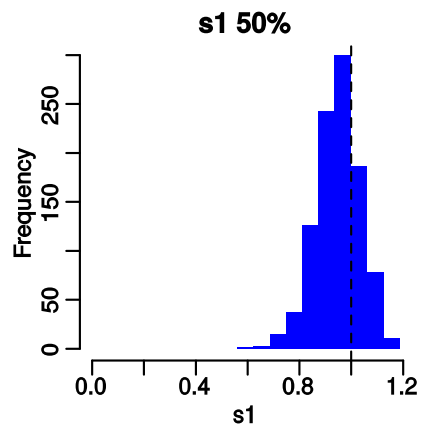
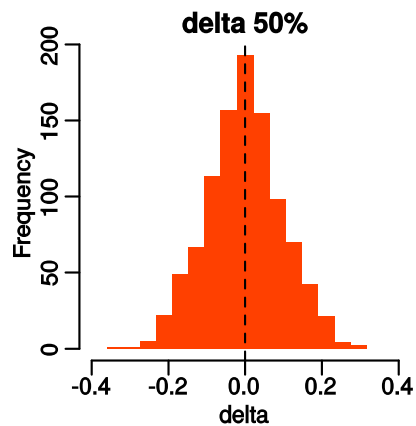
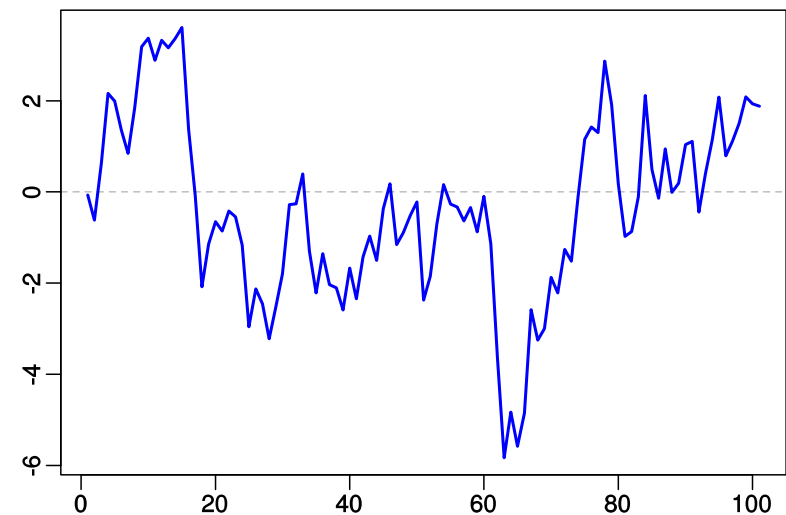
# 状態空間モデルを

「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

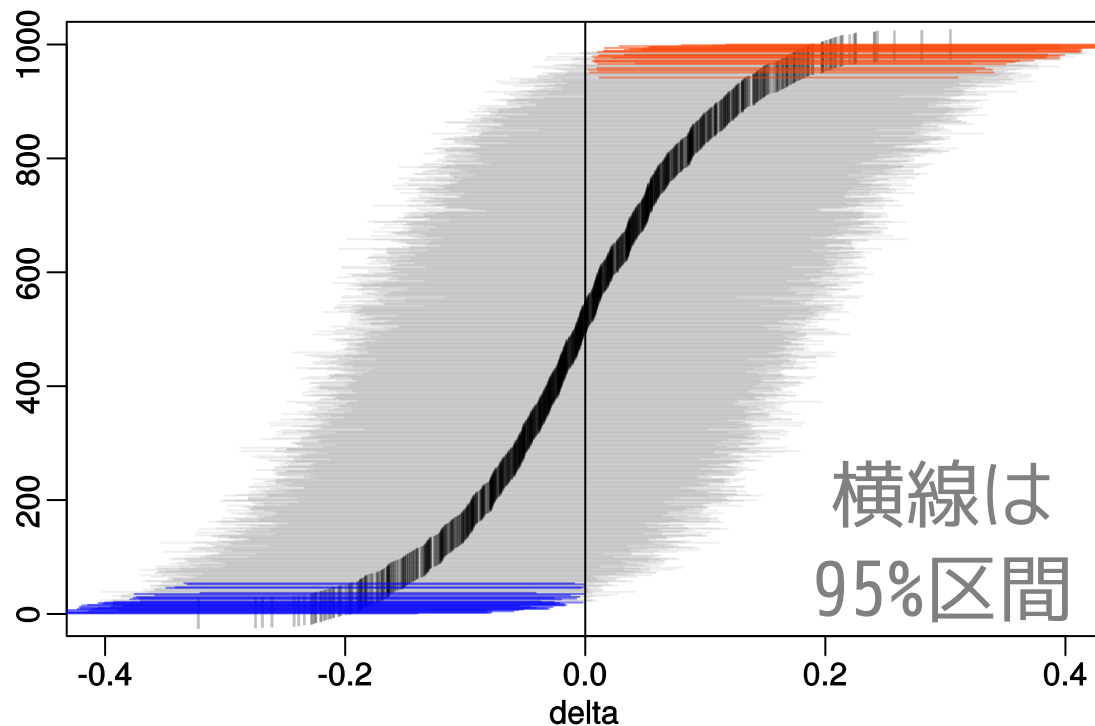
$\sigma_2$  小  
 $\sigma_1$  大  
 $\delta = 0$



# 「傾き」 $\delta$ の事後分布を見る



真の  $\delta$  は 0



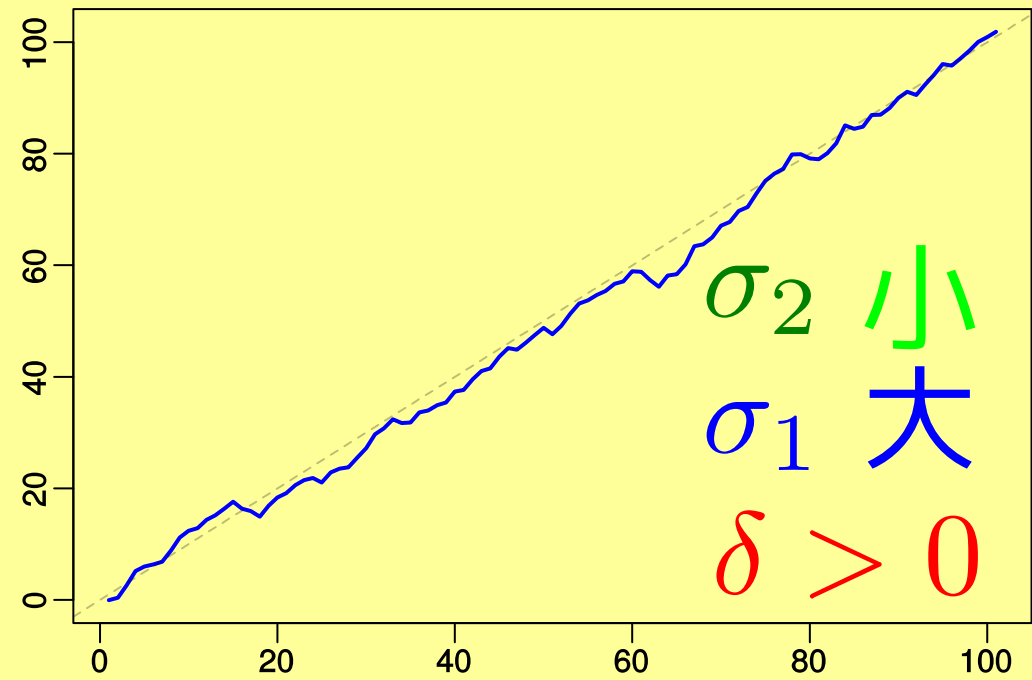
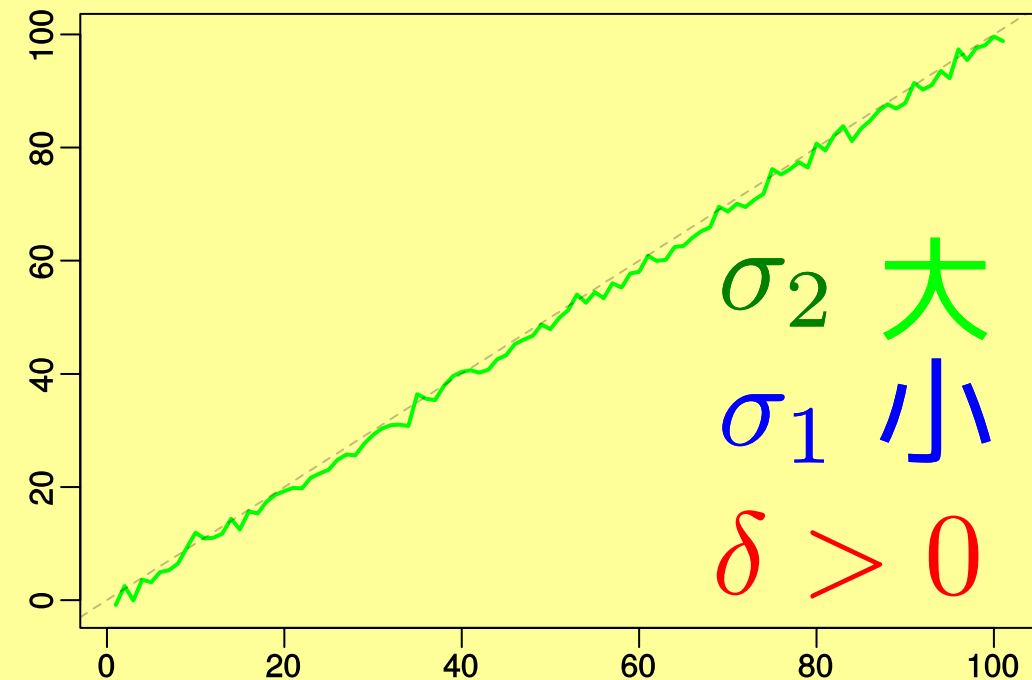
1000回中  
63回ずれた

横線は  
95%区間

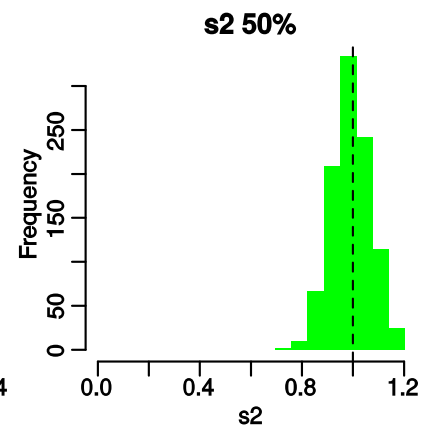
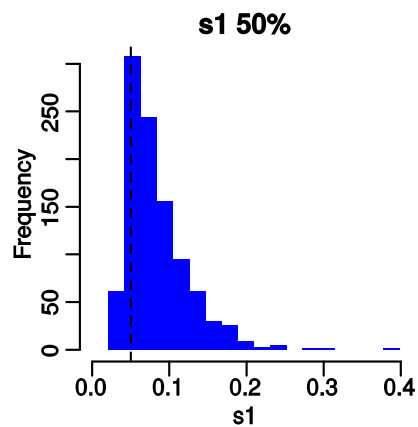
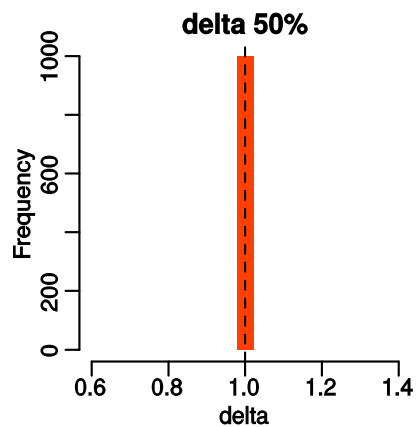
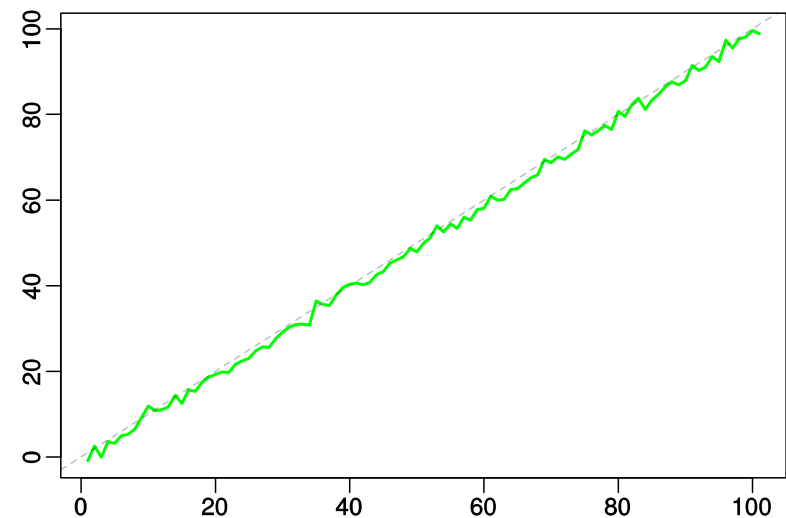
# 状態空間モデルを

「かたむきあり」ランダムウォーク

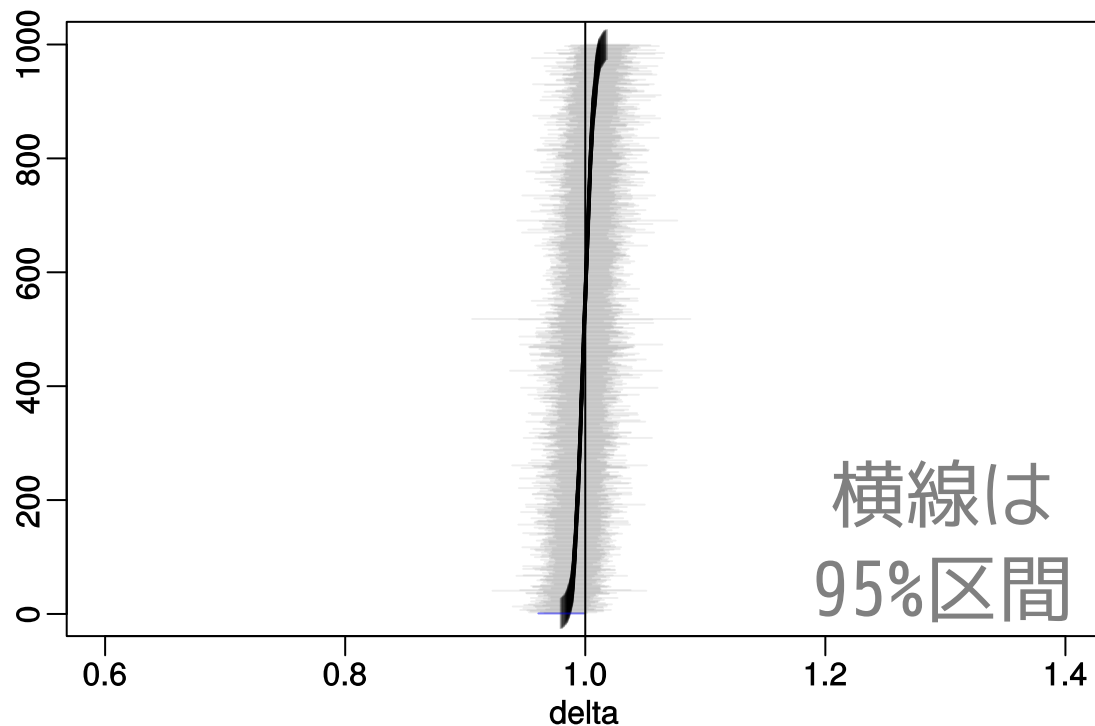
$\delta > 0$   
な架空データにあてはめる



# 「傾き」 $\delta$ の事後分布を見る



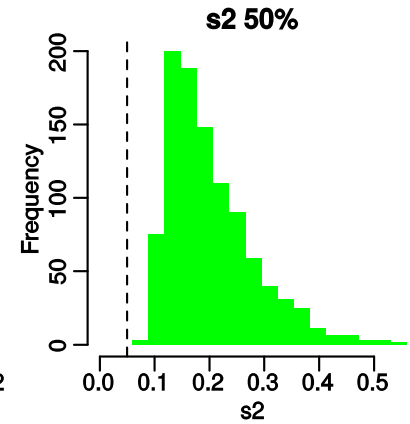
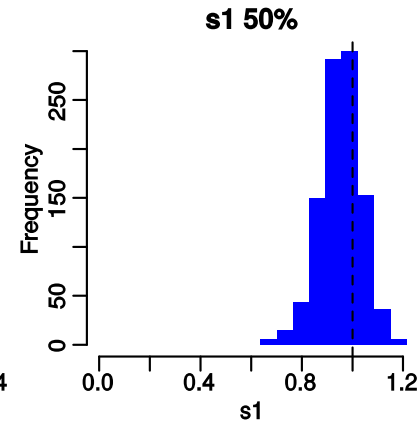
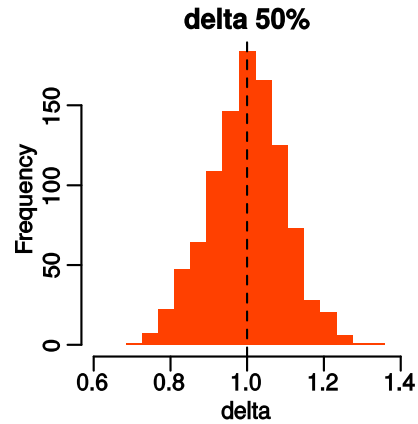
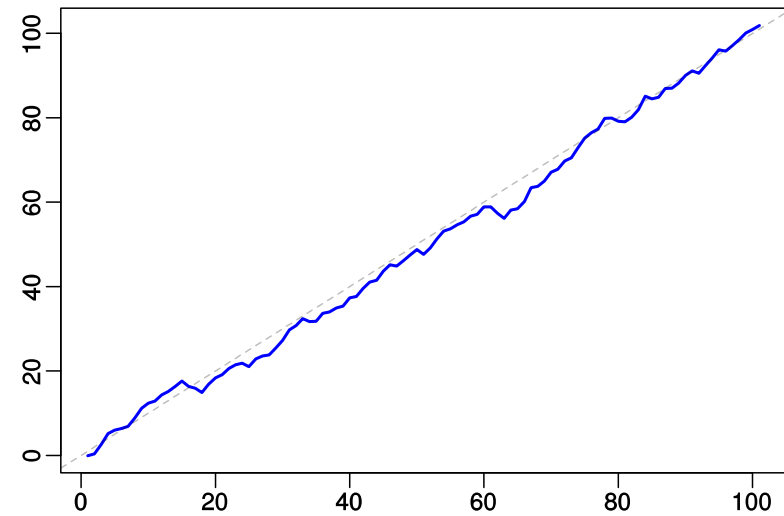
真の $\delta$ は 1



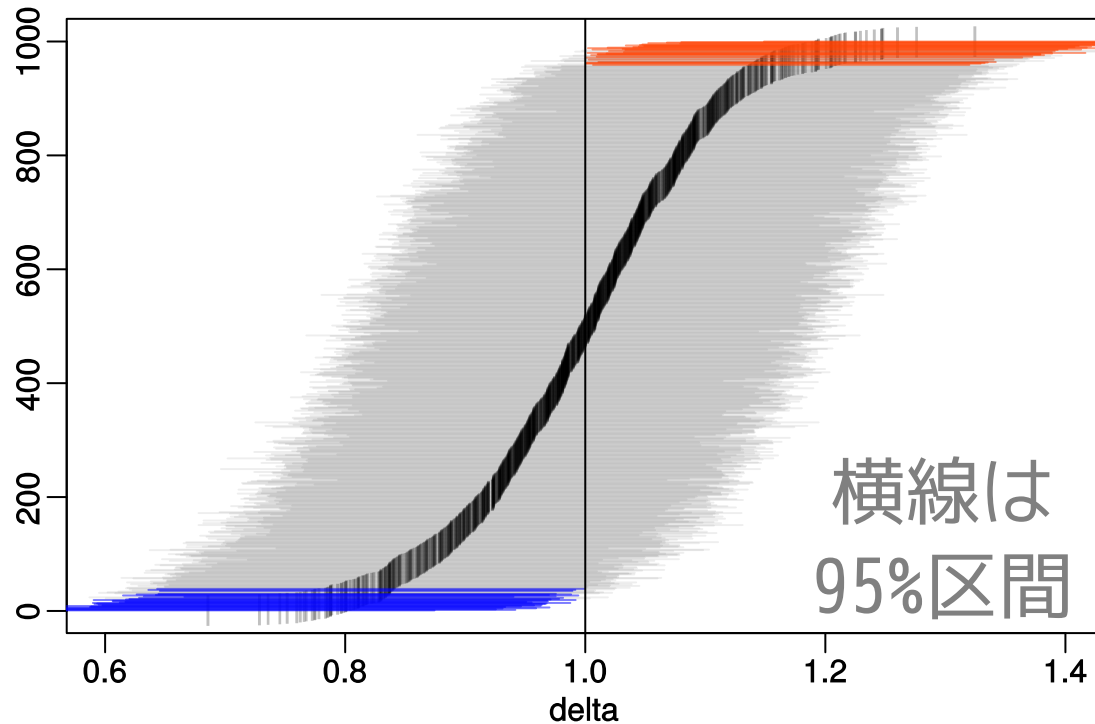
1000回中  
1回ずれた

横線は  
95%区間

# 「傾き」 $\delta$ の事後分布を見る



真の $\delta$ は 1



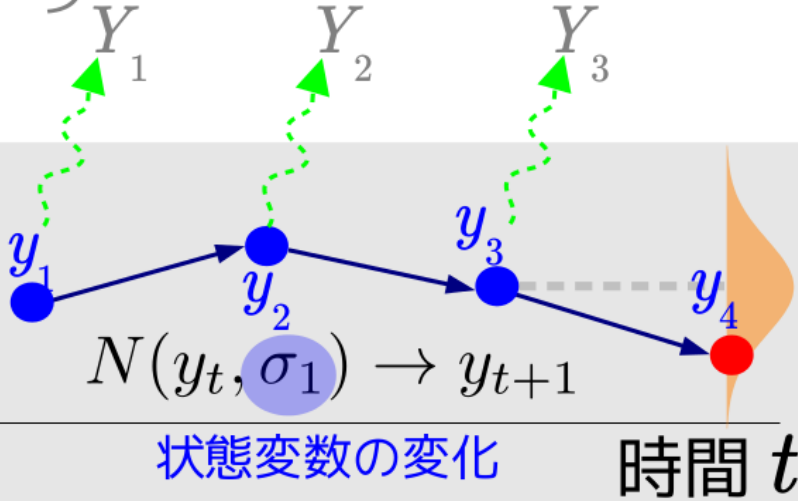
1000回中  
62回ずれた

# とりあえずの結論

観測の誤差 状態空間モデル

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

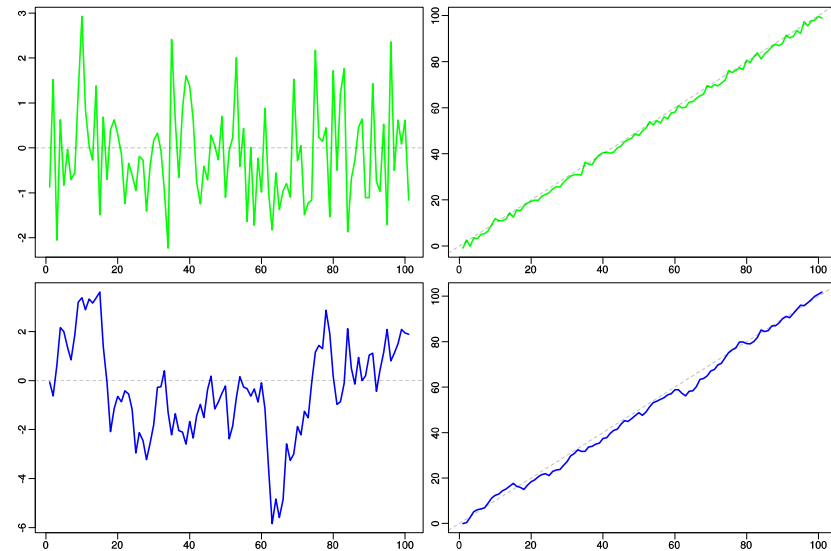
観測データ



観測できない世界 (状態空間)

## 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





# 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
- 欠測と不等間隔
- 時系列「ばらばら解析」やめよう
- 「うたがわしい回帰」への対策

階層ベイズモデル!

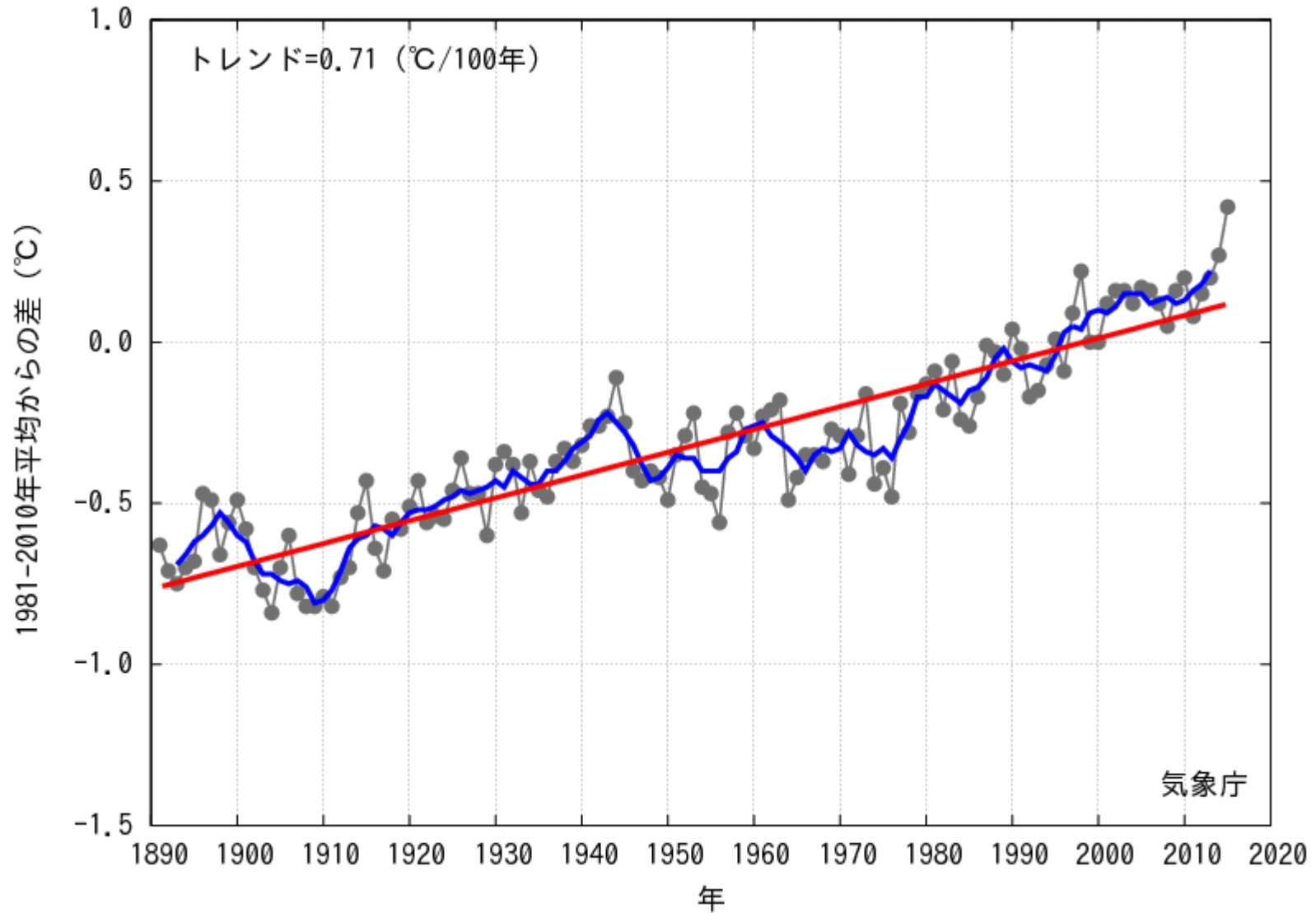
# 状態空間モデルを使う利点

「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

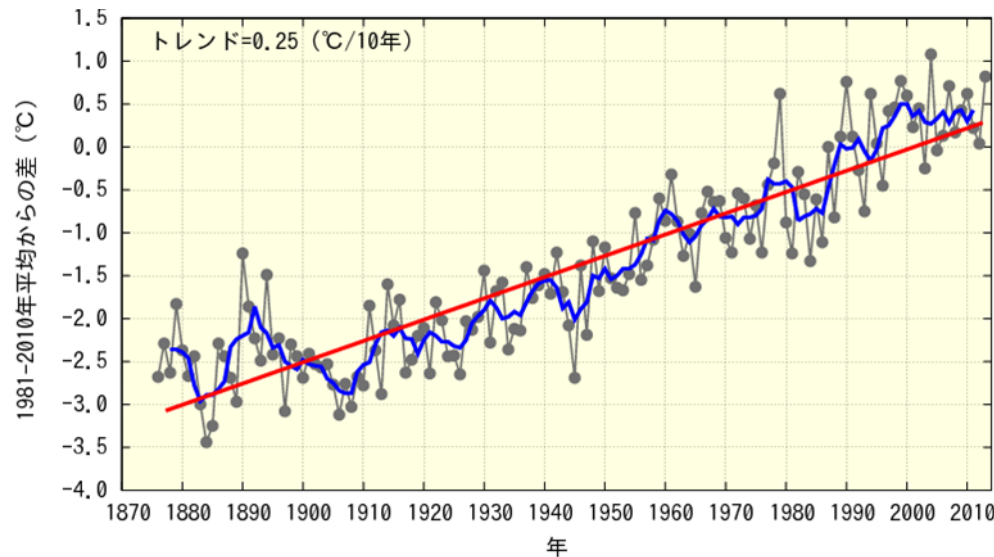
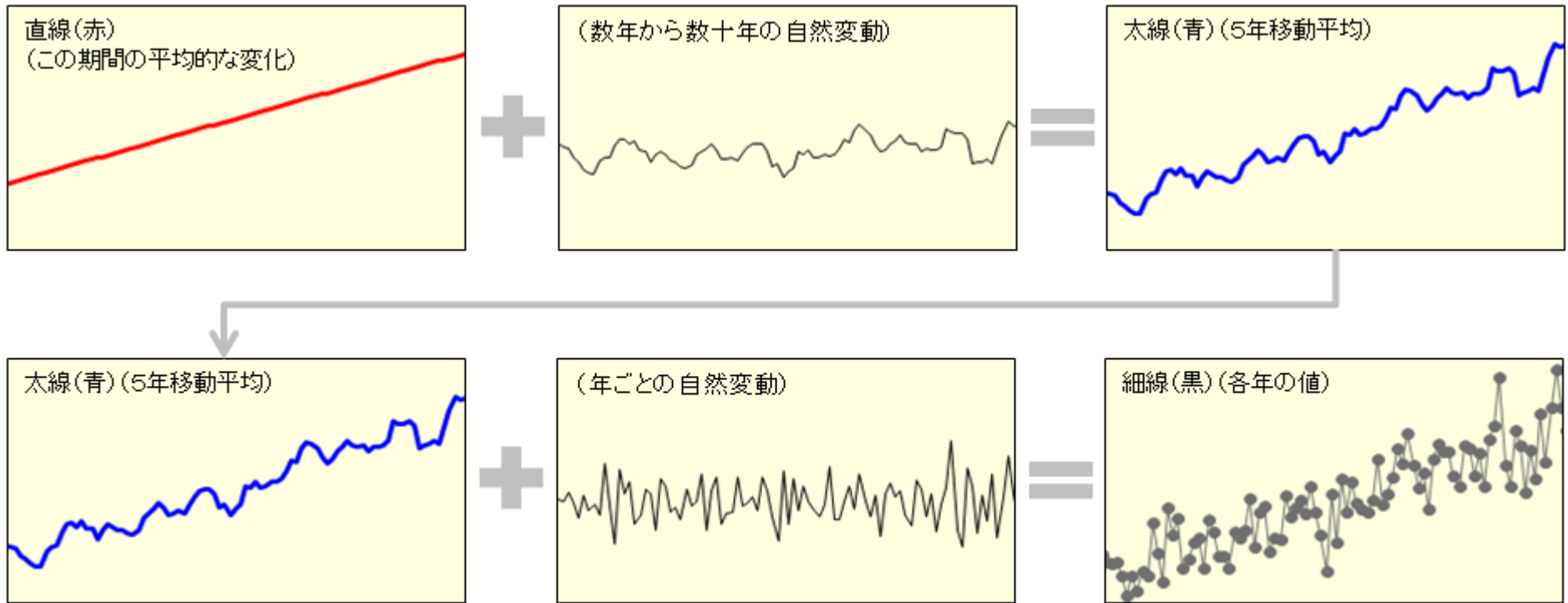
世界の年平均気温偏差



[http://www.data.jma.go.jp/cpdinfo/temp/an\\_wld.html](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

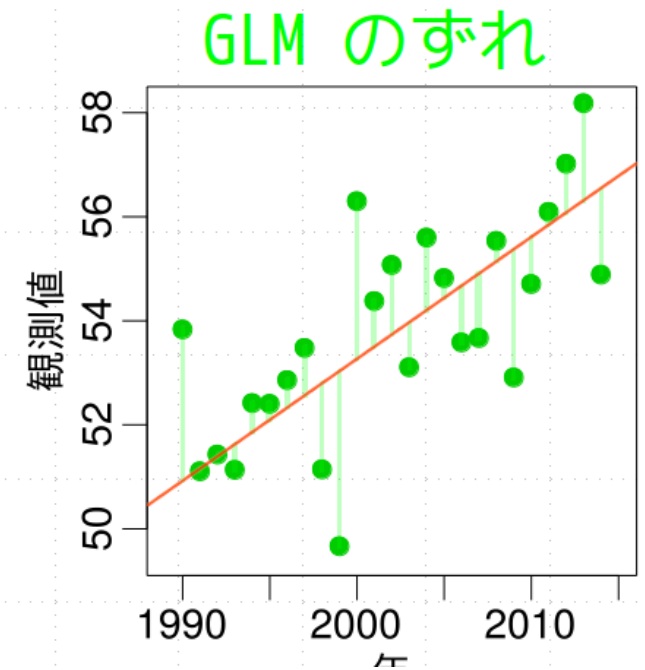
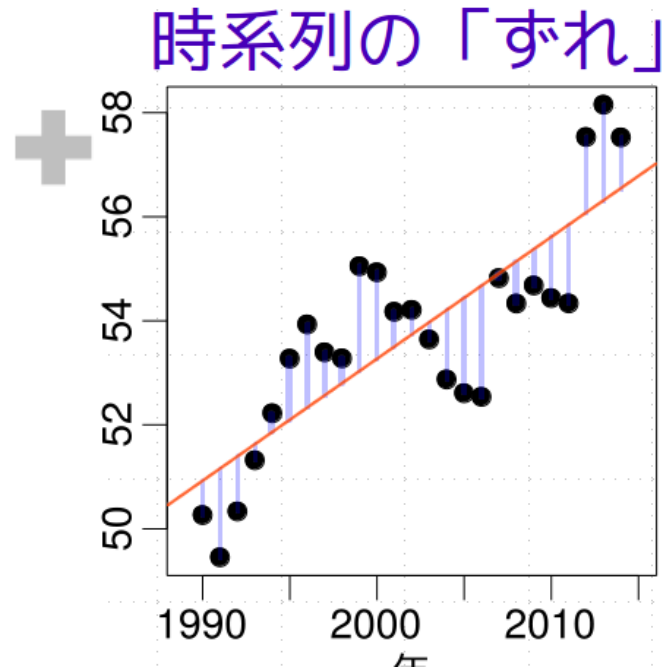
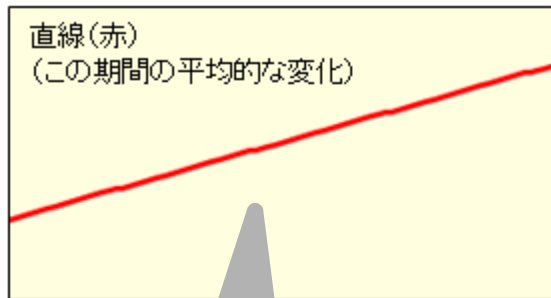
<http://www.data.jma.go.jp/cpdinfo/temp/trend.html>



気象庁  
ばらばら  
メソッド?

# 気象庁ばらばらメソッド何がまずいか？

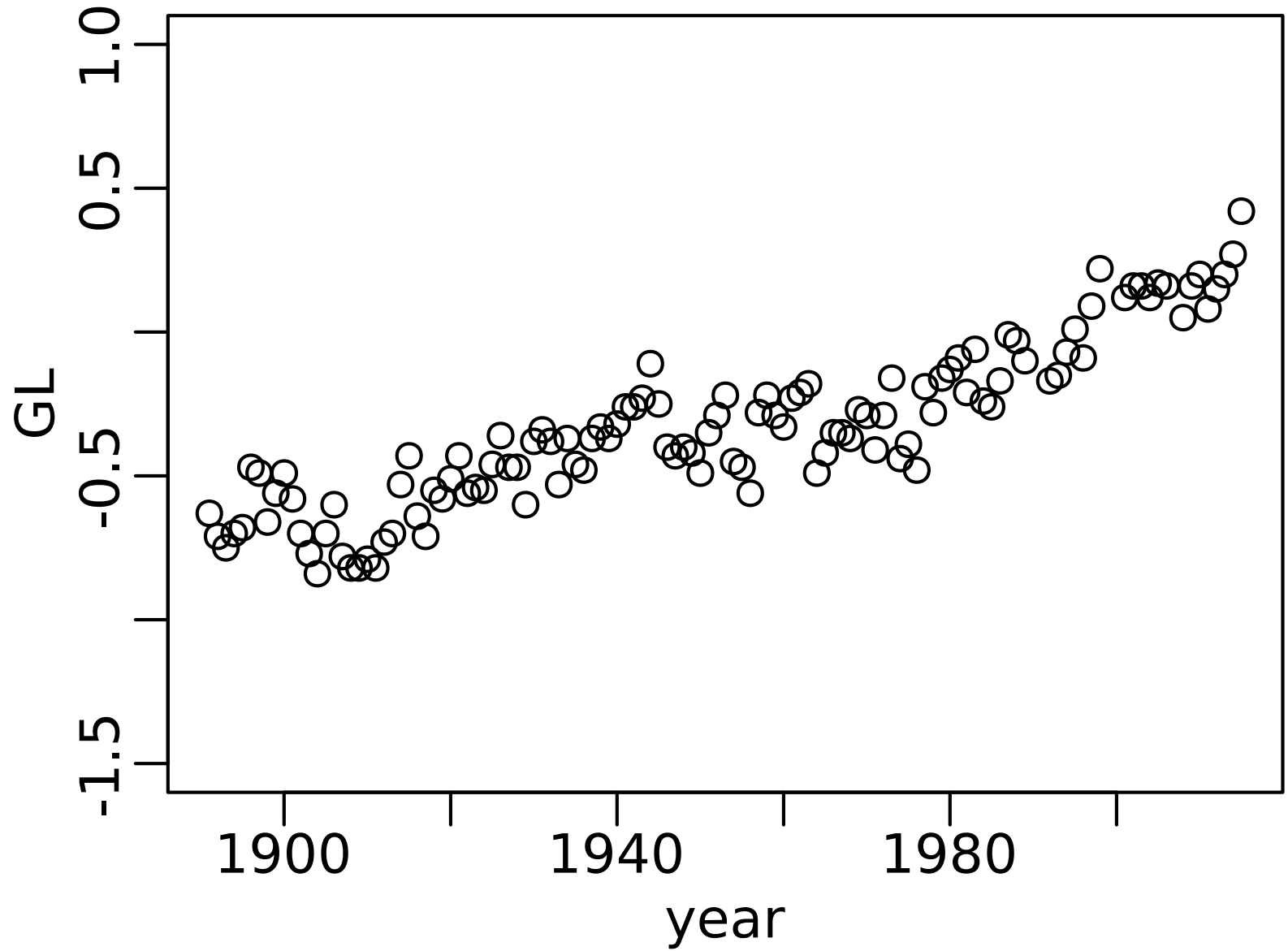
<http://www.data.jma.go.jp/cpdinfo/temp/trend.html>



いきなり直線回帰!

(時系列データなんだから…)

# 公開データをダウンロード

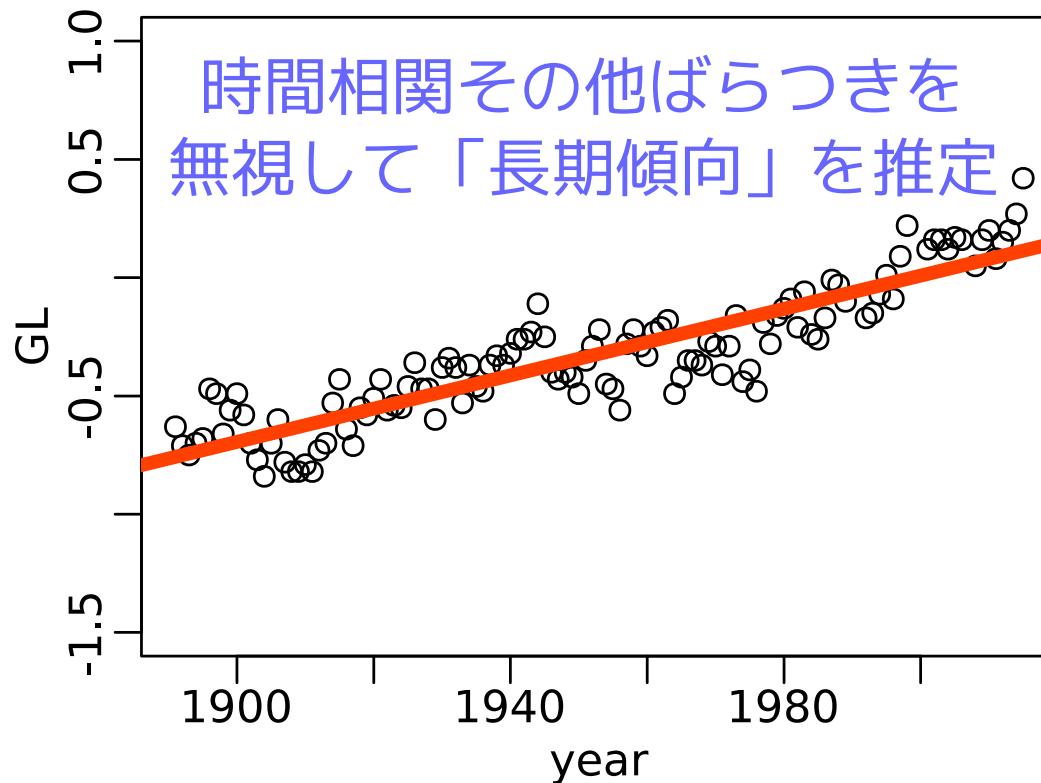


# 「とりあえず、直線回帰」の危険性

```
> 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



確率 1京ぶんの 2?

100年  
あたり  
0.70°C

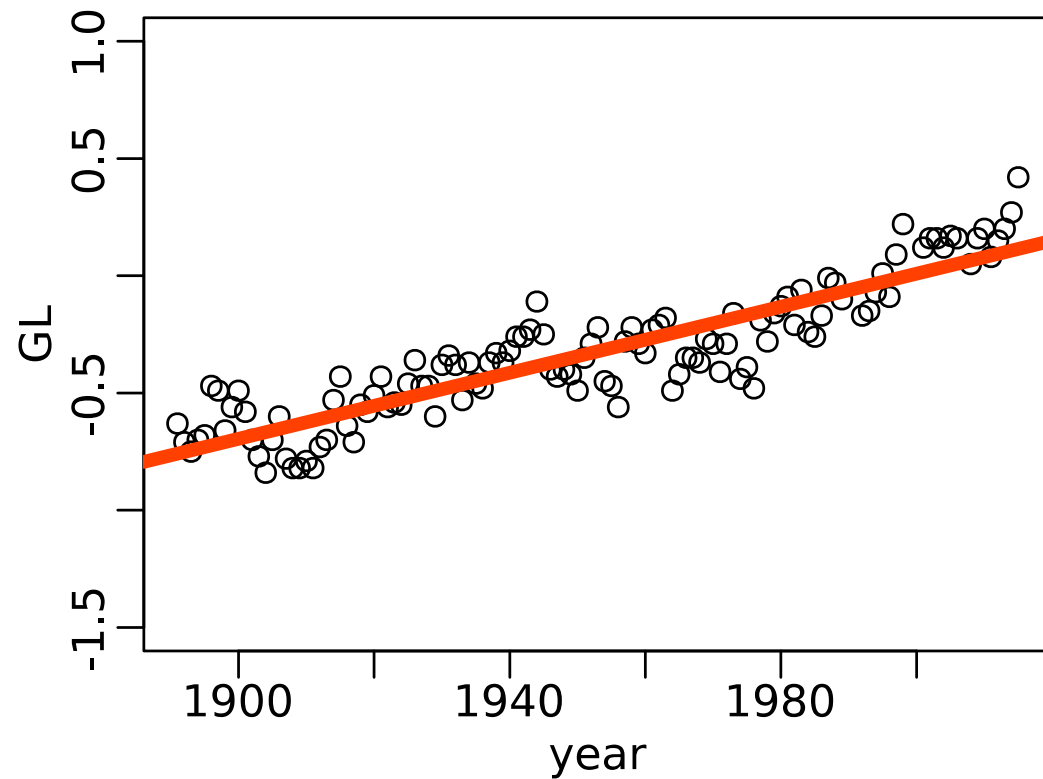
# 直線あてはめ (GLM) が予測した「温暖化」

```
> 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°C

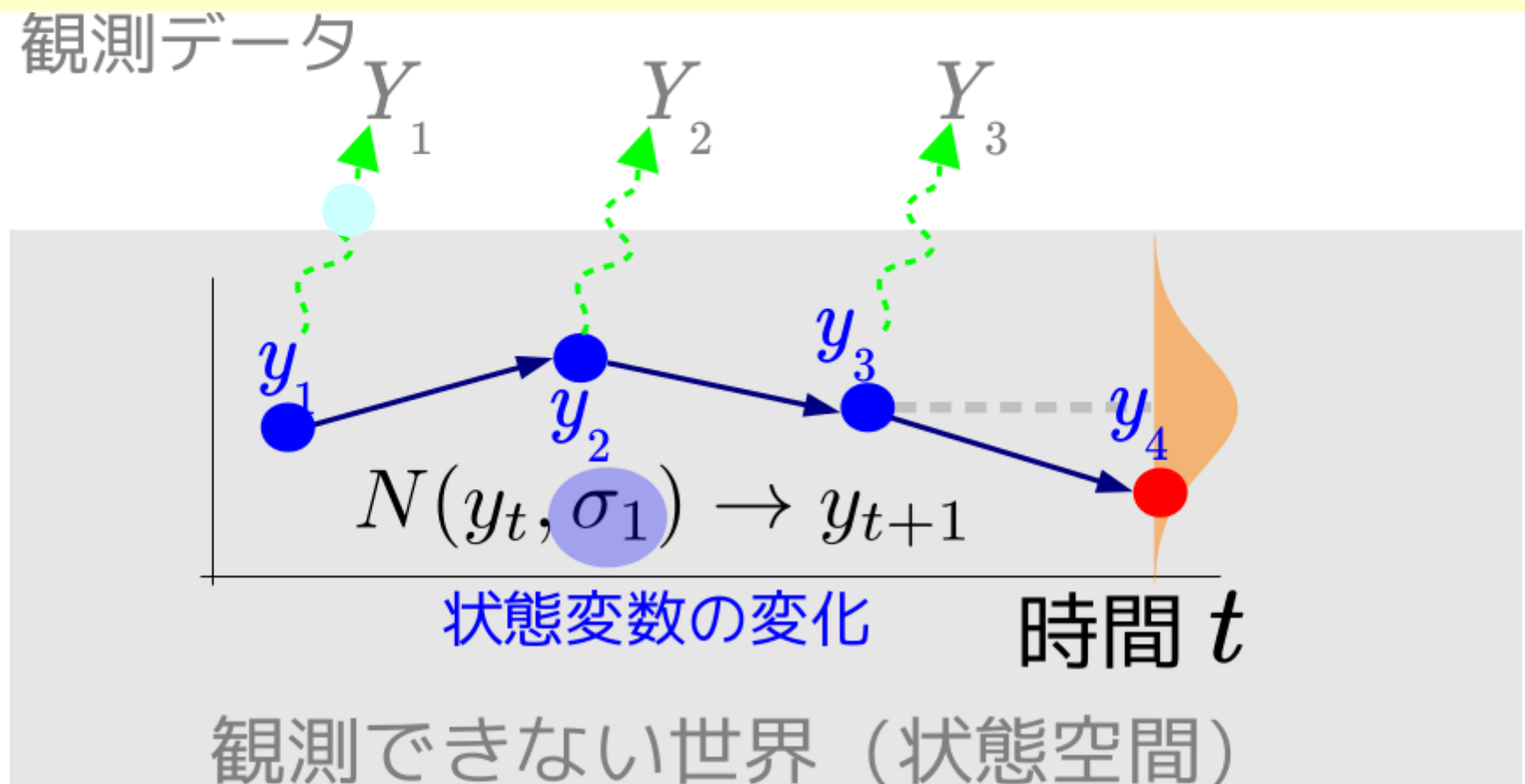




# 状態空間モデル：すべてを同時に推定

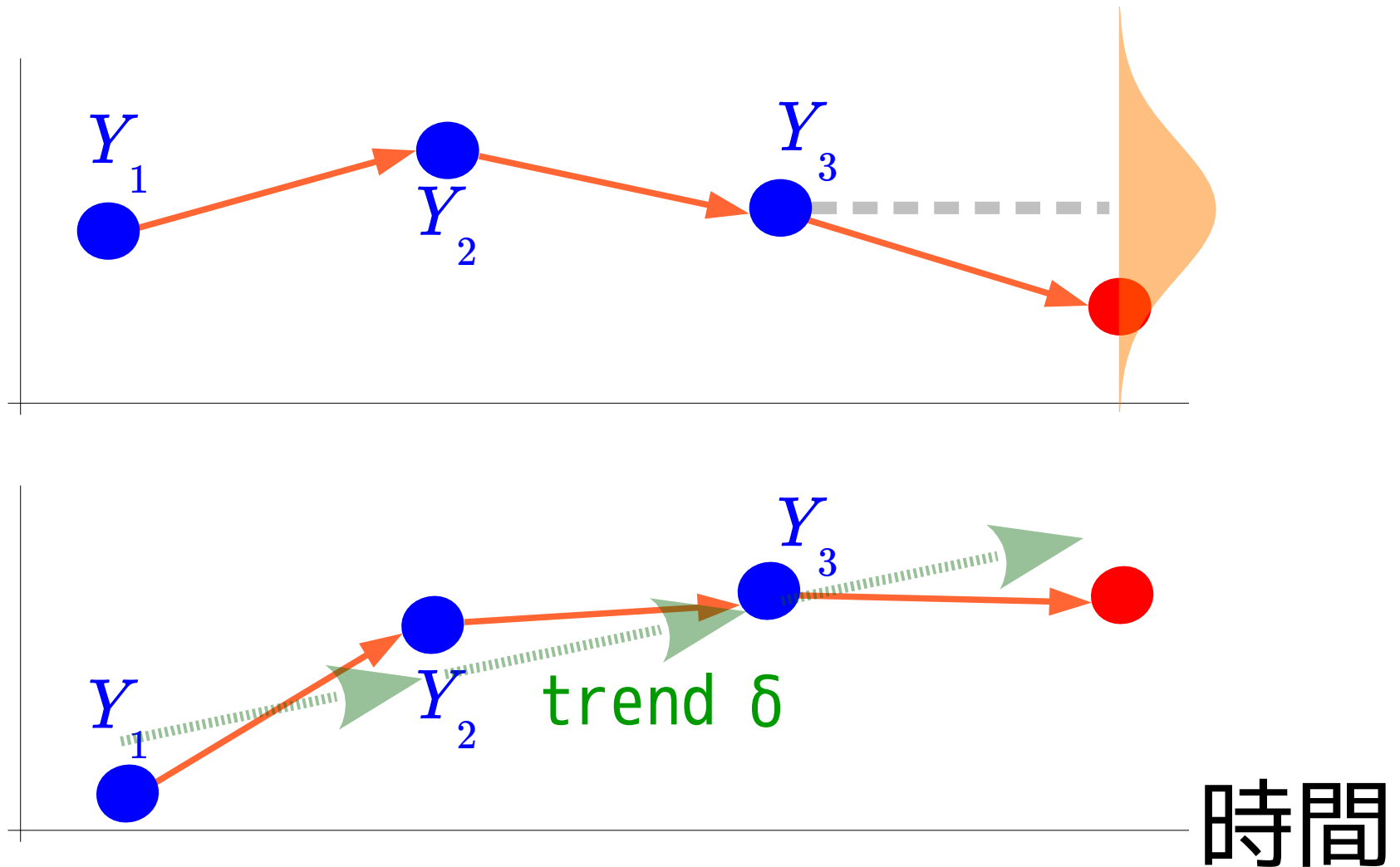
ランダムウォーク+各年独立なノイズ

観測データ



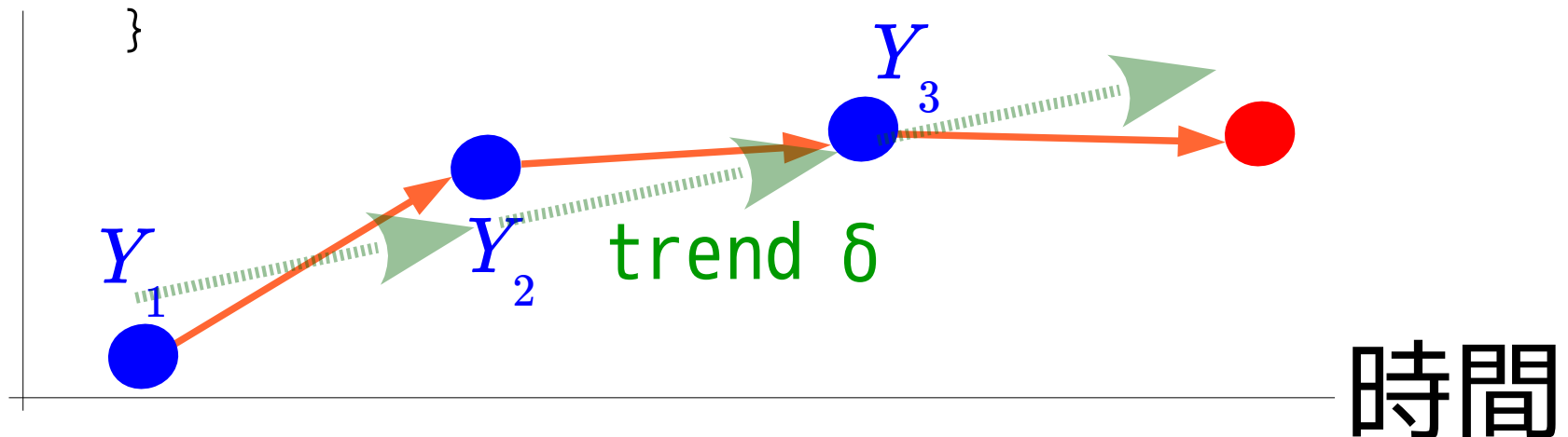
# 状態空間モデル：すべてを同時に推定

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)
}
```



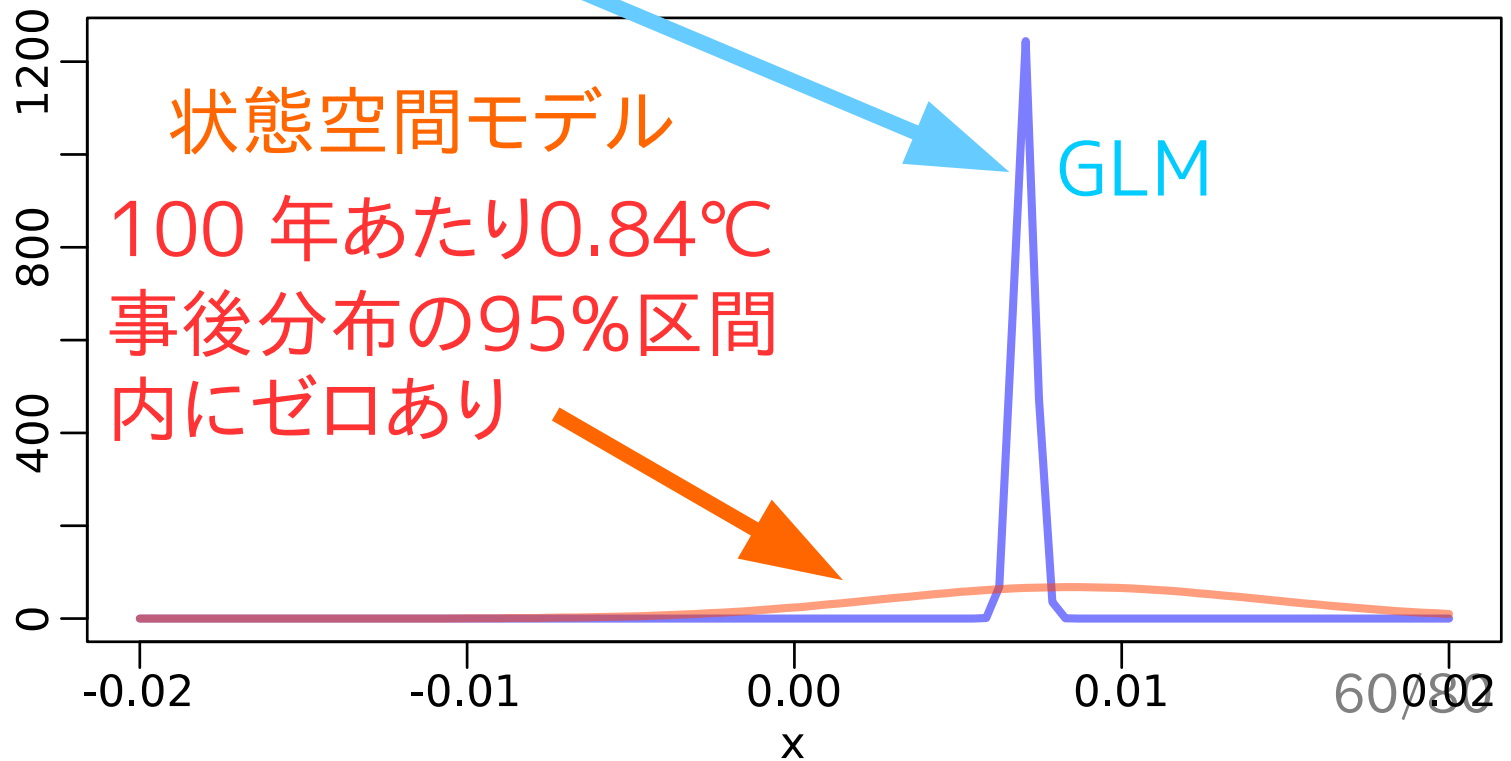
# 状態空間モデルが予測した「温暖化」

```
> 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°C

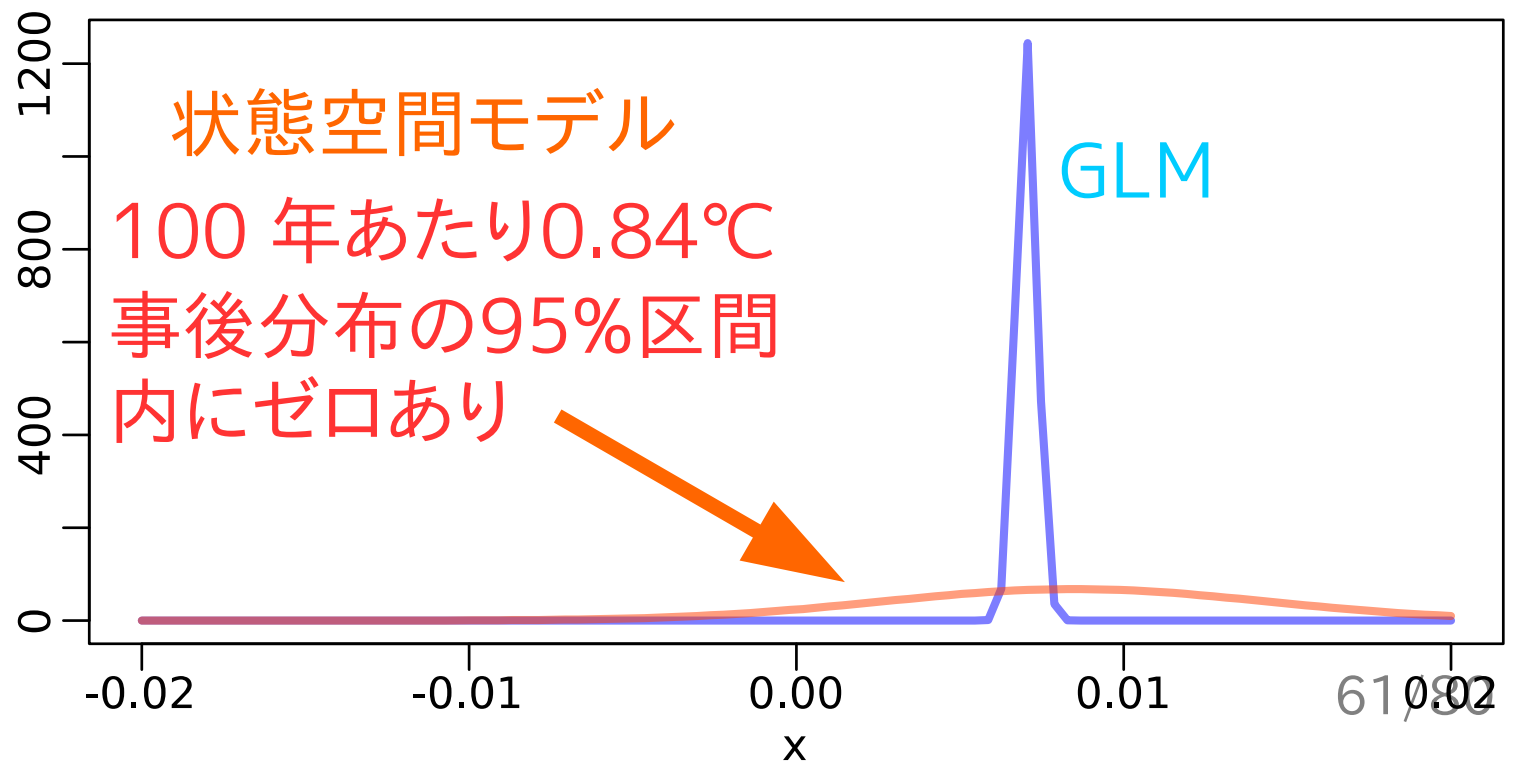


観測値間に相関あり →

実質的な

サンプルサイズが小さくなる

100年  
あたり  
0.70°C



疑わしい回帰  
spurious regression

時系列どうしの回帰

time series  $Y \sim$  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

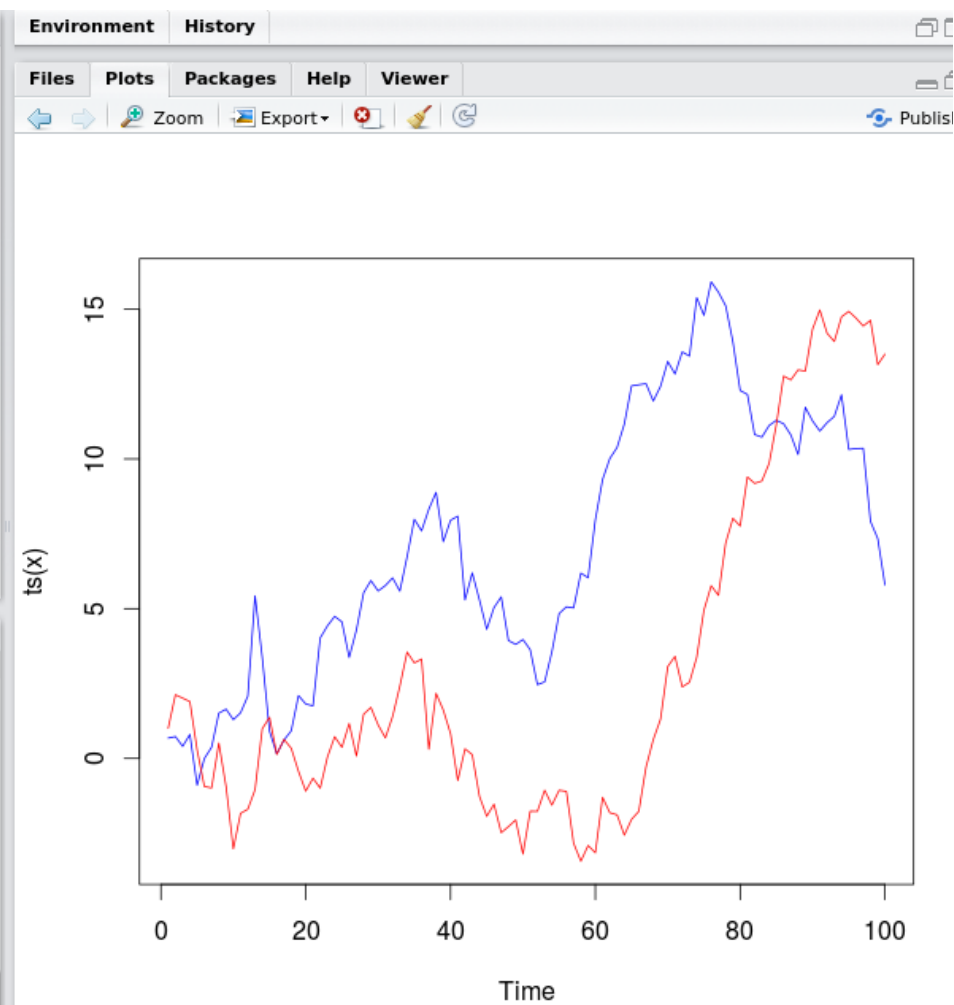
- GLM:  $Y(t) \sim t$  とか  $Y(t) \sim X(t)$
- 段階的解析: 観測値の四則演算
- 「残差」の再解析
- 「対応」の無視 – 再測は時系列

# 「見せかけの回帰」 spurious regression

```
spurious_regression.R x
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
```

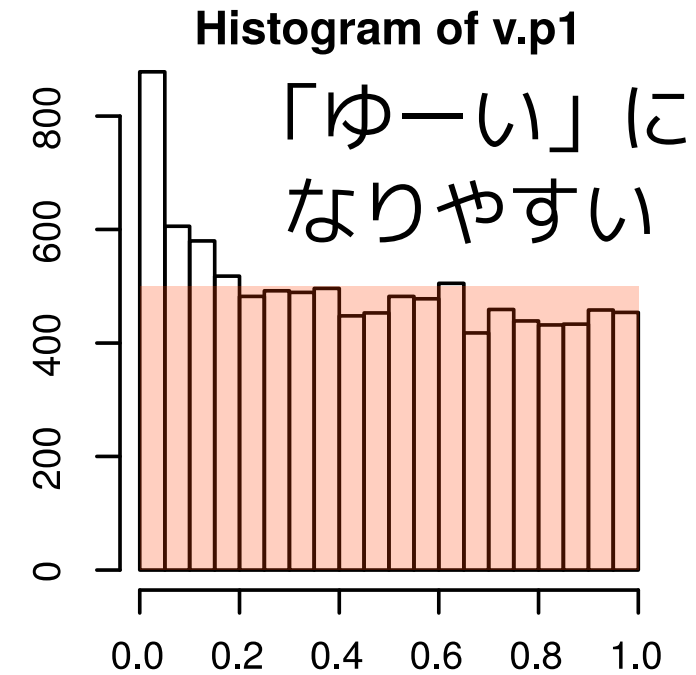
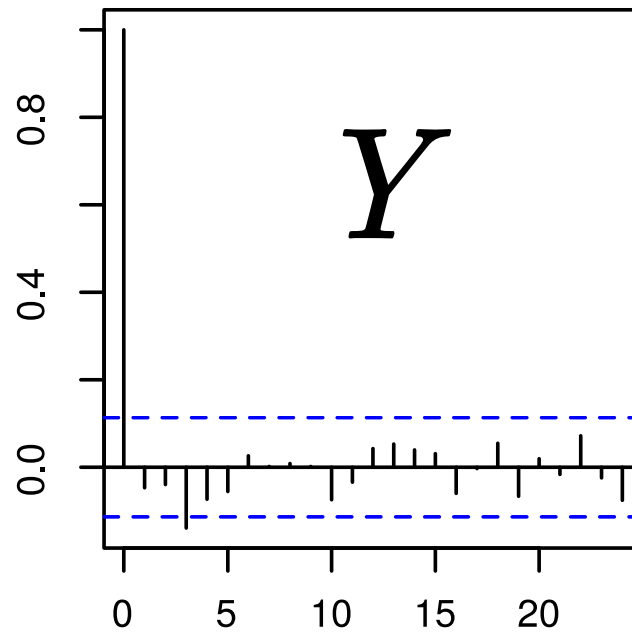
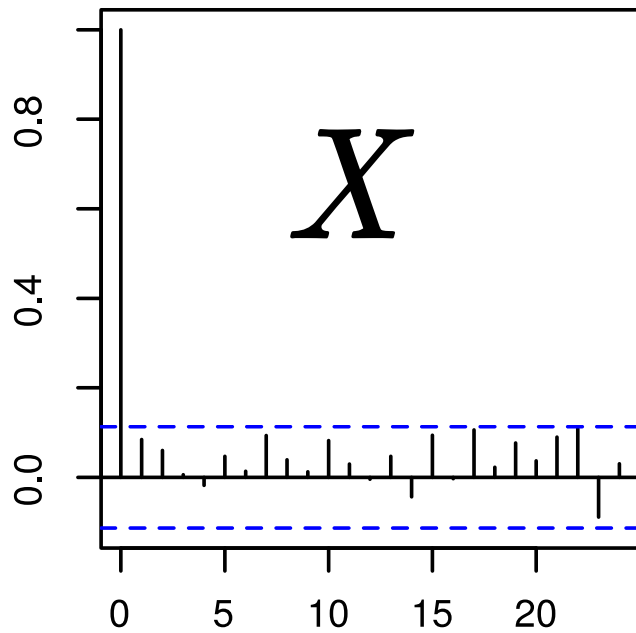


ちょっとだけ実演してみます



# ノイズの大きな時系列にうもれたワナ？

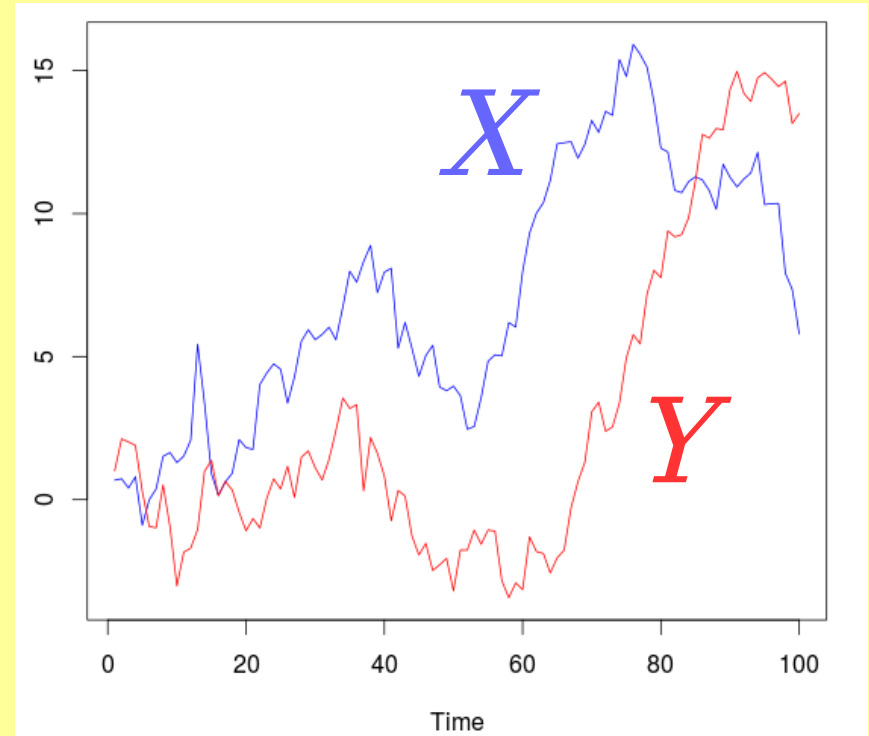
## 時間的自己相関のない時系列？



しかし  $\text{glm}(Y \sim X)$  とすると…

$$Y \sim X$$

疑わしい回帰  
spurious  
regression



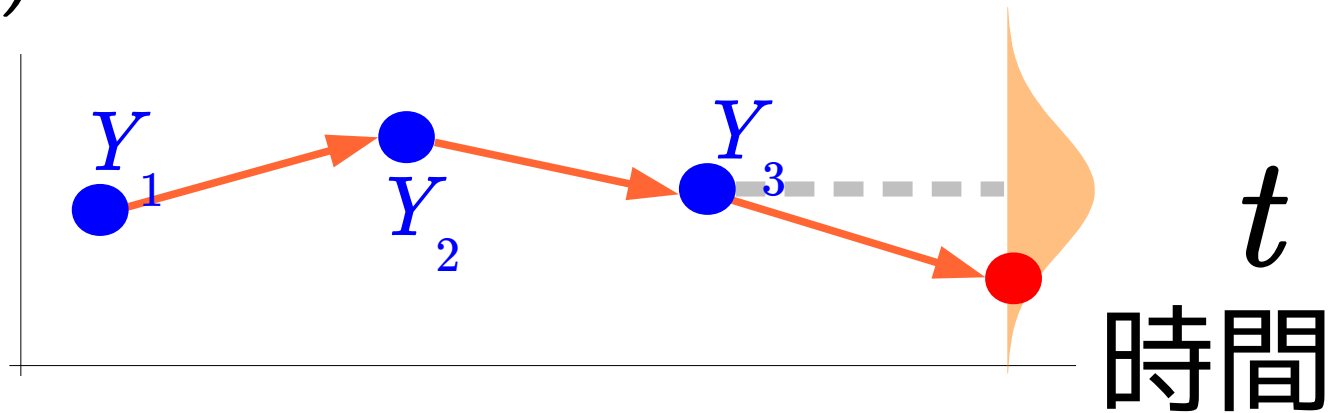
この問題も

状態空間モデル (SSM)で

解決できないだろうか?

# 二変量のランダムウォーク モデルを作れないか?

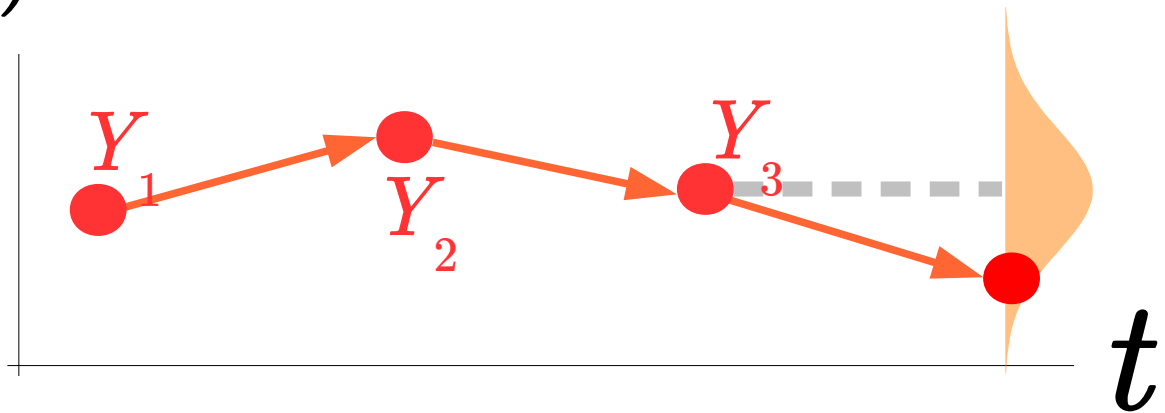
$$Y_{t+1} \sim N(Y_t, s_y)$$



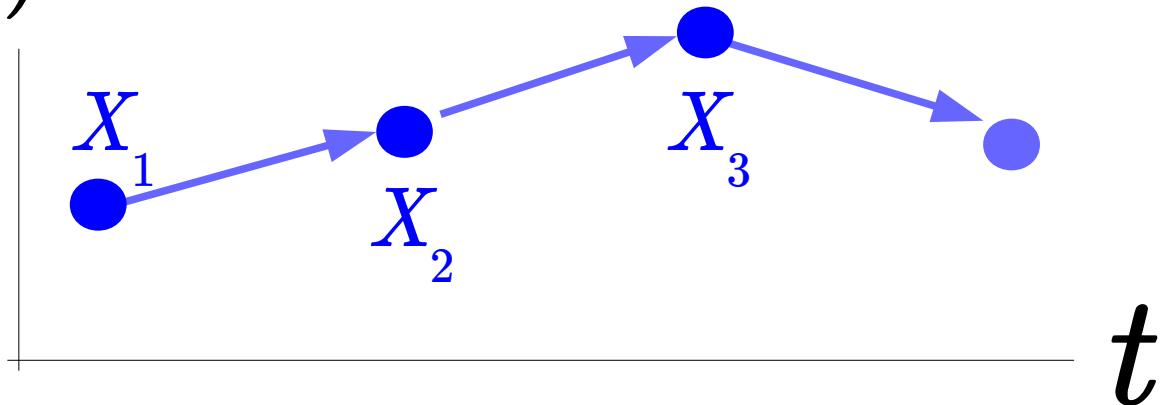
# 二変量のランダムウォーク

$Y_t$  と  $X_t$  は独立

$$Y_{t+1} \sim N(Y_t, s_y)$$



$$X_{t+1} \sim N(X_t, s_x)$$



# 二変量のランダムウォーク

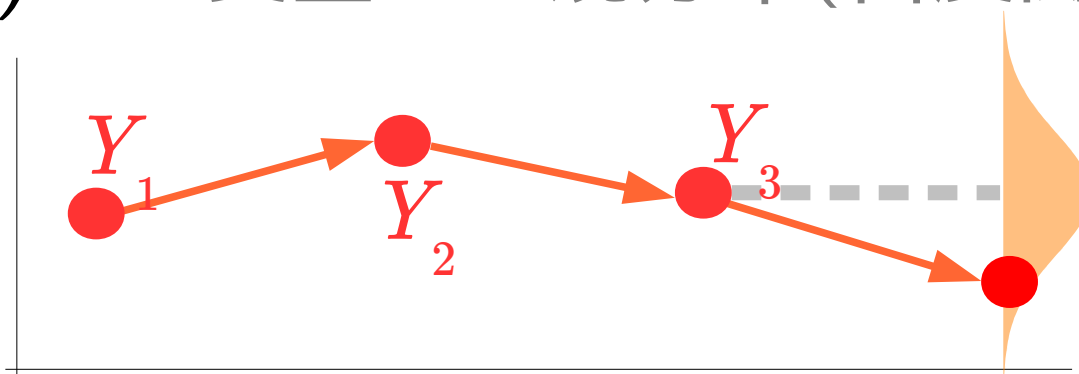
$Y_t$  と  $X_t$  は独立

$$Y_{t+1} \sim N(Y_t, s_y)$$

$$X_{t+1} \sim N(X_t, s_x)$$

このあたりで  
何とかならないか?

$Y_{t+1} \sim N(Y_t, s_y)$  一変量の正規分布(密度関数)



## 二変量の正規分布(密度関数)

**Bivariate case**

In the 2-dimensional nonsingular case ( $k = \text{rank}(\Sigma) = 2$ ), the probability density function of a vector  $[X \ Y]'$  is:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right]\right)$$

where  $\rho$  is the correlation between  $X$  and  $Y$  and where  $\sigma_X > 0$  and  $\sigma_Y > 0$ . In this case,

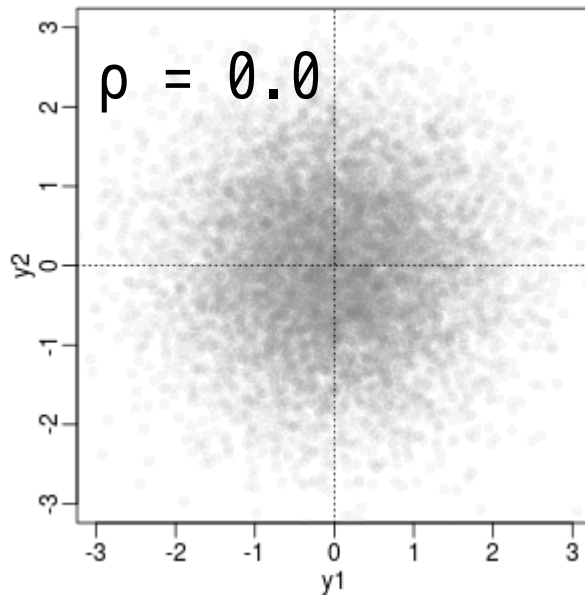
$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

相関係数  $\rho$

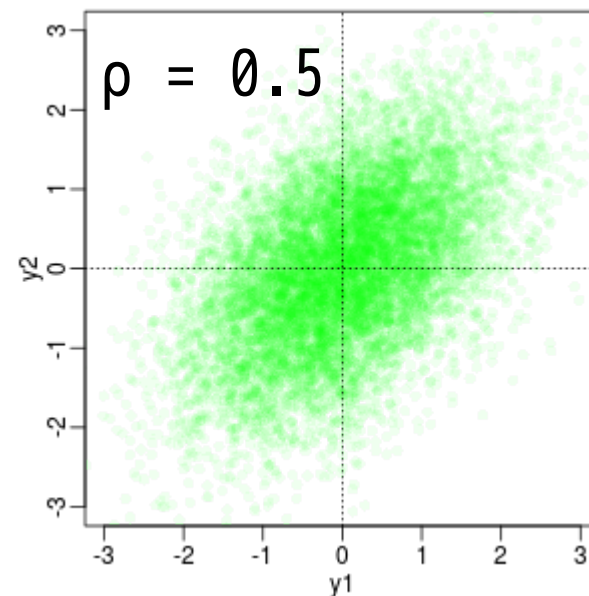
分散共分散行列

[https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution](https://en.wikipedia.org/wiki/Multivariate_normal_distribution)

無相関



正の相関



## 二変量の正規分布(密度関数)

### Bivariate case

In the 2-dimensional nonsingular case ( $k = \text{rank}(\Sigma) = 2$ ), the probability density function of a vector  $[X \ Y]'$  is:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right]\right)$$

where  $\rho$  is the correlation between  $X$  and  $Y$  and where  $\sigma_X > 0$  and  $\sigma_Y > 0$ . In this case,

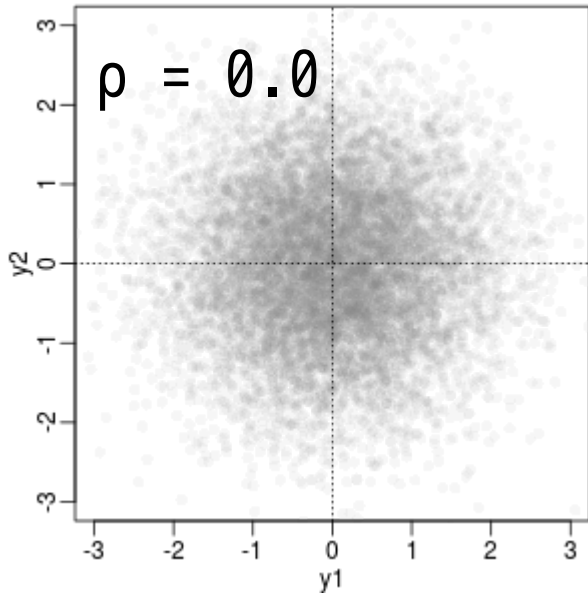
$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

相関係数  $\rho$

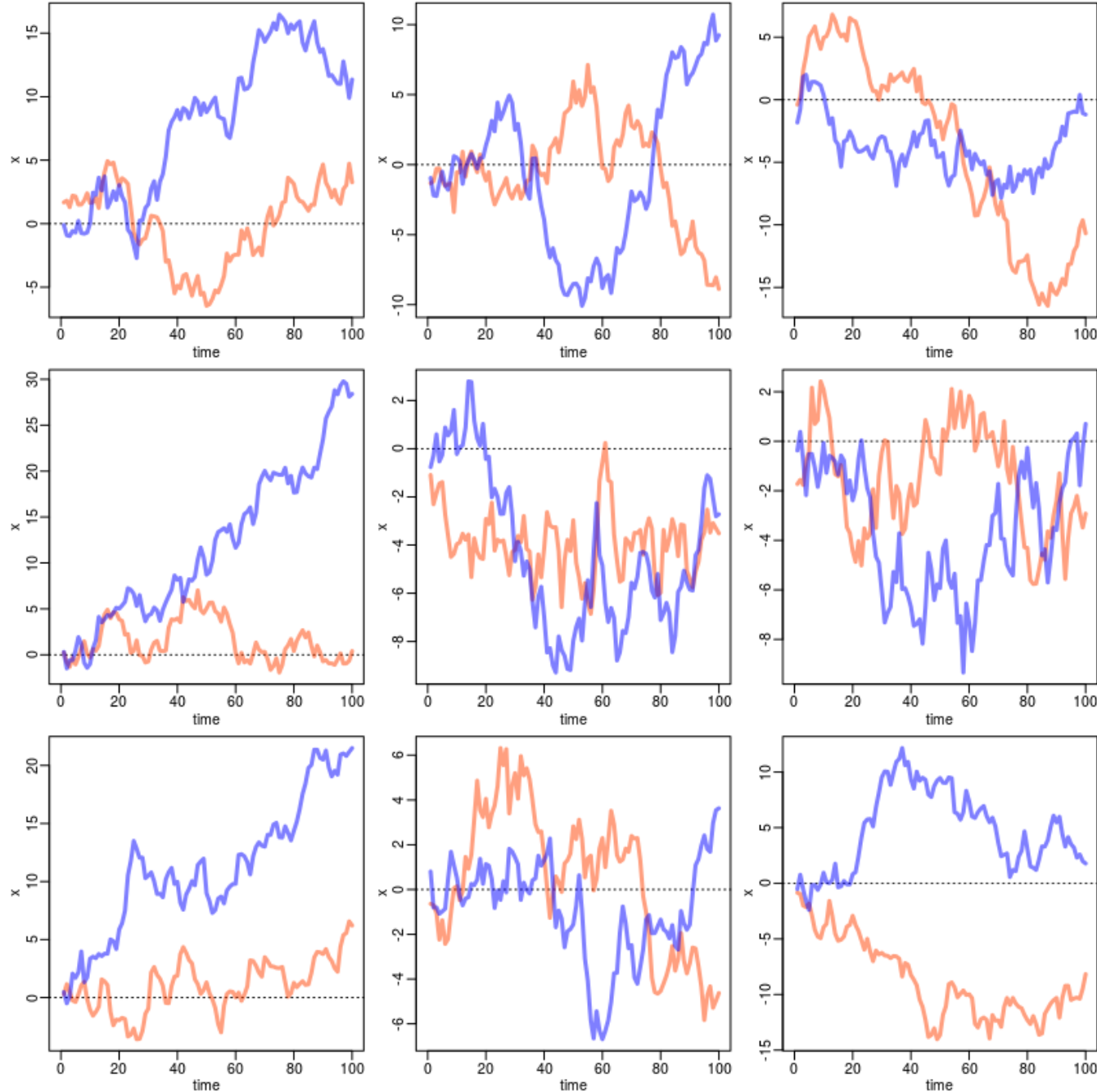
分散共分散行列

[https://en.wikipedia.org/wiki/Multivariate\\_normal\\_distribution](https://en.wikipedia.org/wiki/Multivariate_normal_distribution)

# 二変量正規分布とランダムウォーク 例1

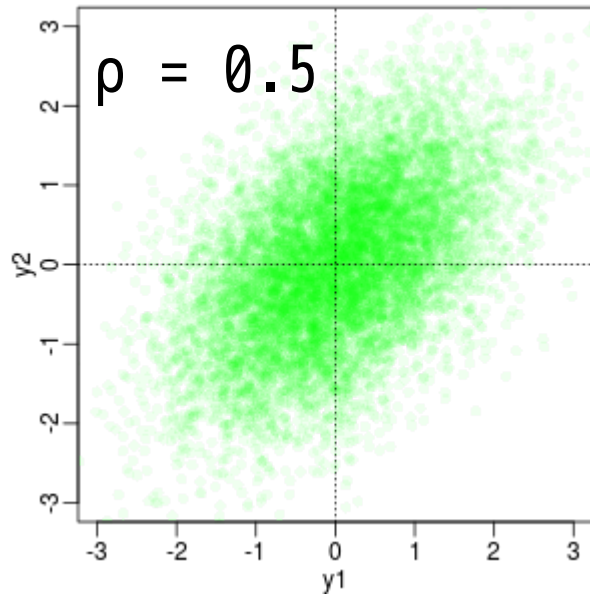


無相関





# 二変量正規分布とランダムウォーク 例2

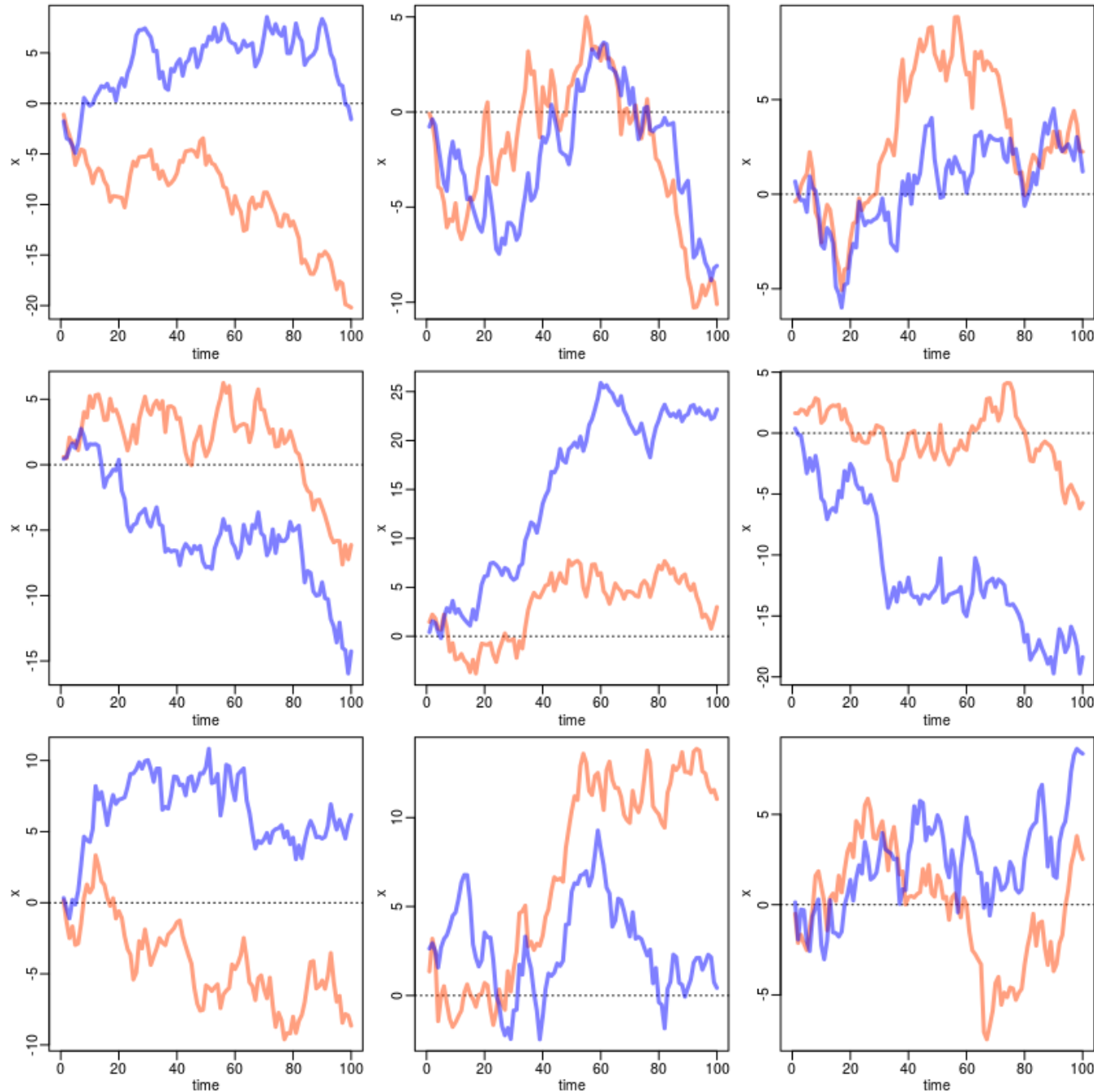


正の相関

時間があれば

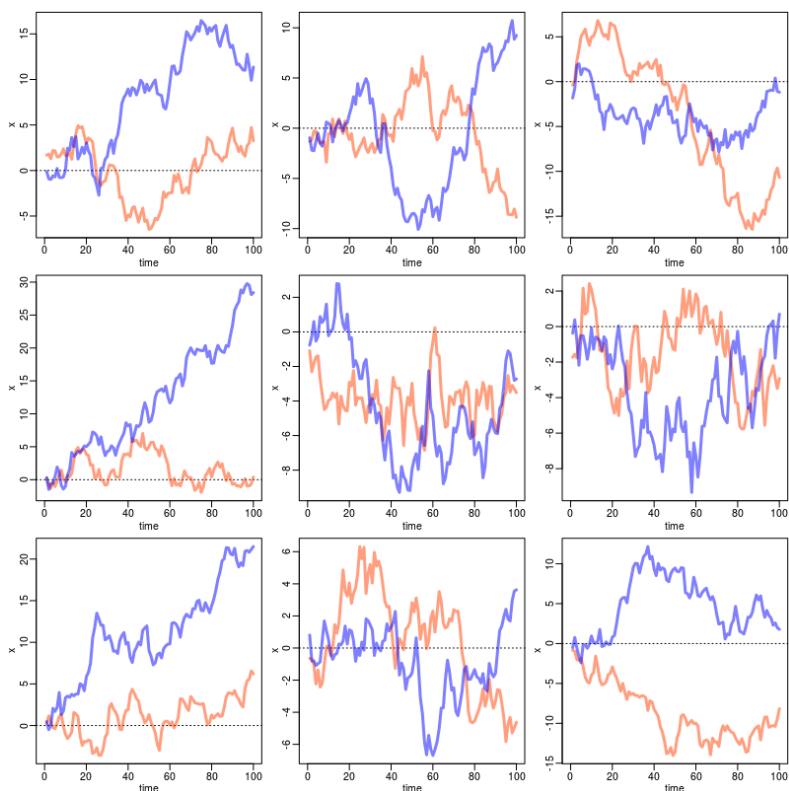
demo

`sample_rvar.R`

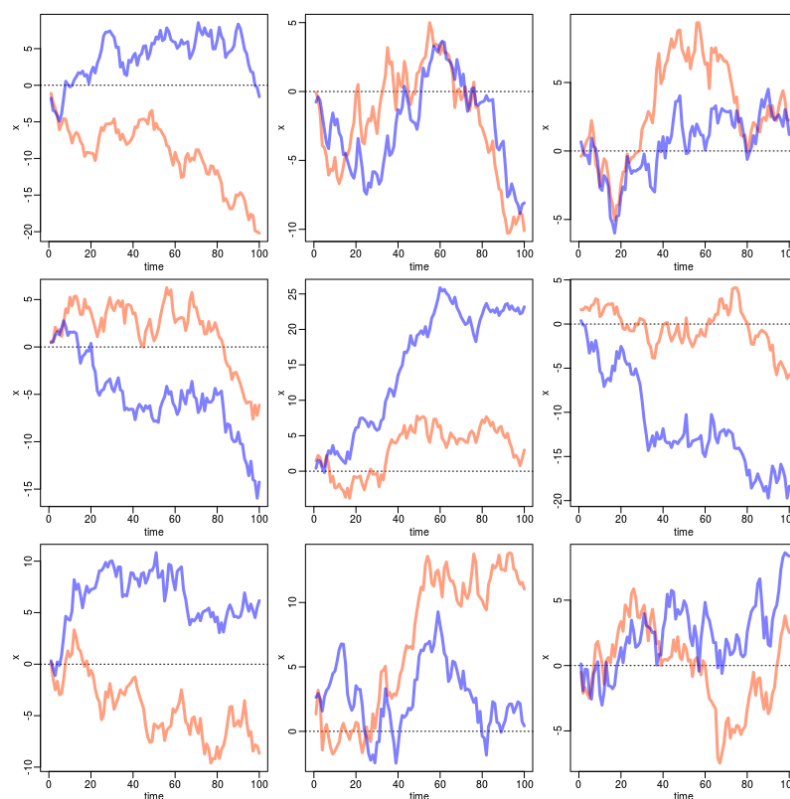


# 二変量正規分布とランダムウォーク

$X$  と  $Y$  の相関係数  $\rho$  を推定できるのか?



$$\rho = 0.0$$



$$\rho = 0.5$$

# 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] ~ dnorm(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

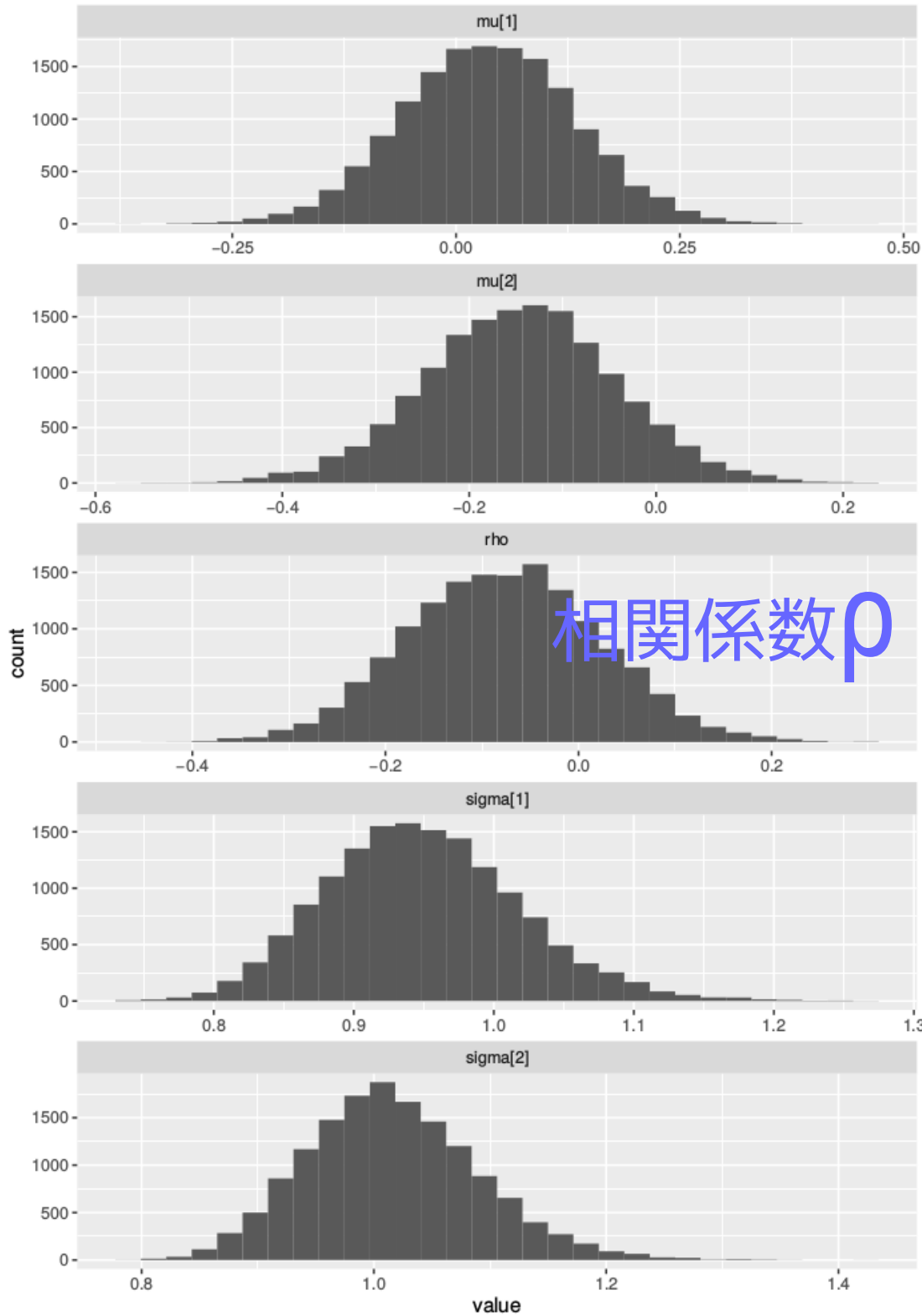
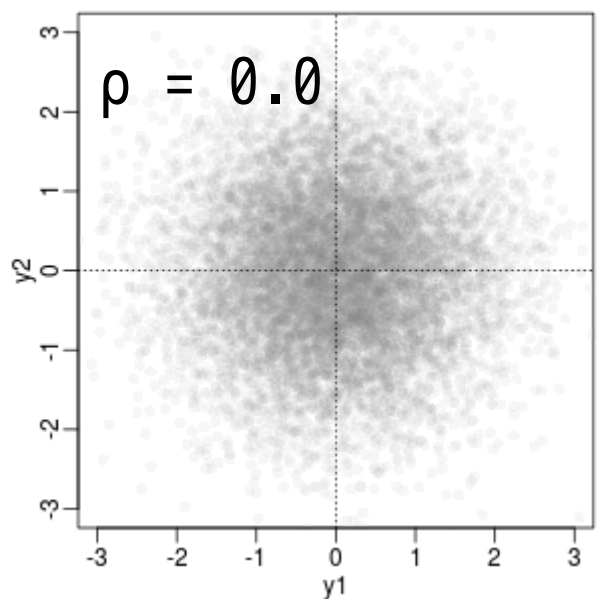
```
3 chains, each with 5200 iterations (first 200 discarded)
n.sims = 15000 iterations saved
      mean      sd    2.5%    25%    50%    75%  97.5%  Rhat  n.eff
mu[1]  -0.122  0.110  -0.342  -0.195  -0.120  -0.048  0.090  1.001  6000
mu[2]  -0.157  0.100  -0.355  -0.224  -0.157  -0.091  0.041  1.002  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

図示すると……

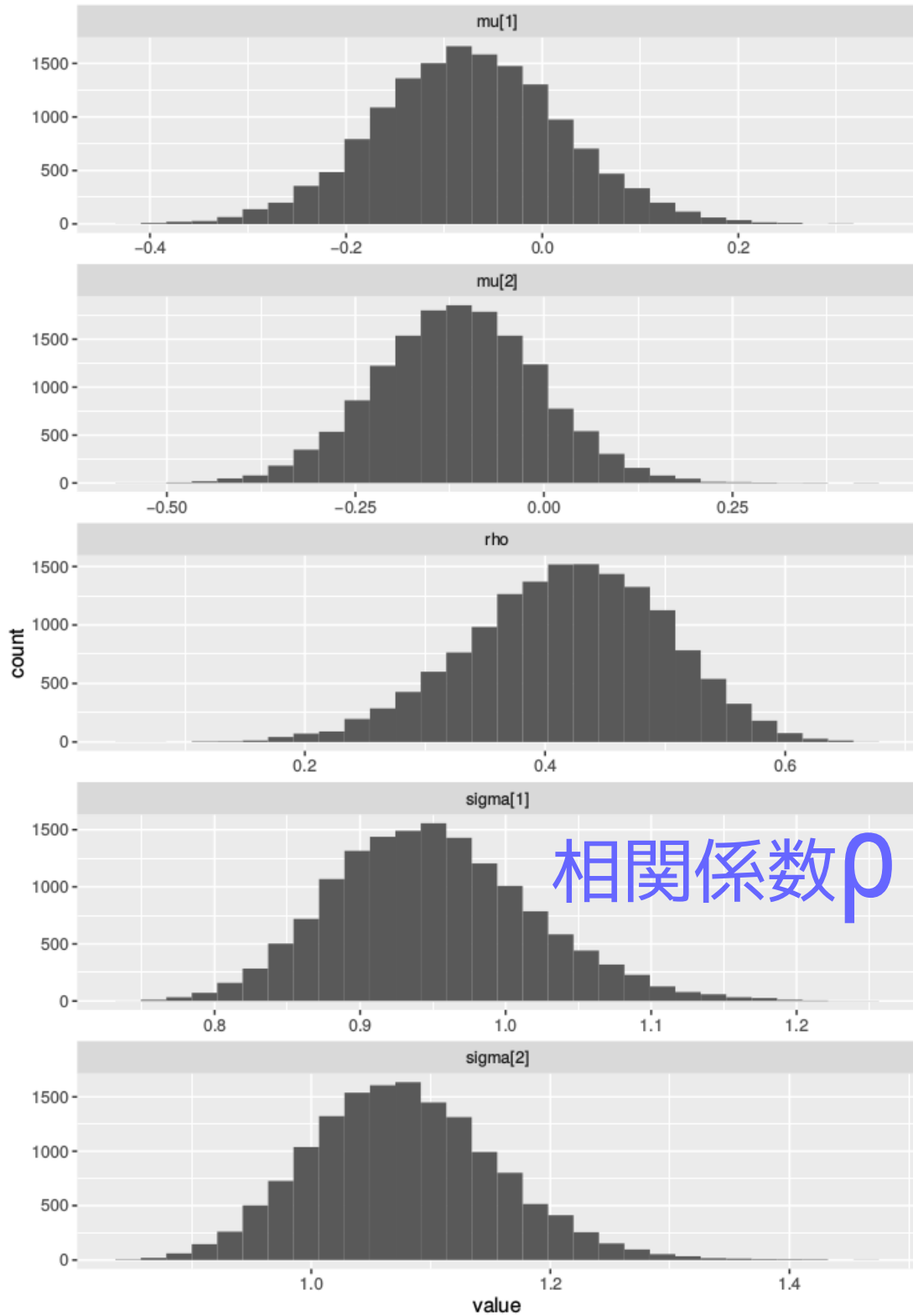
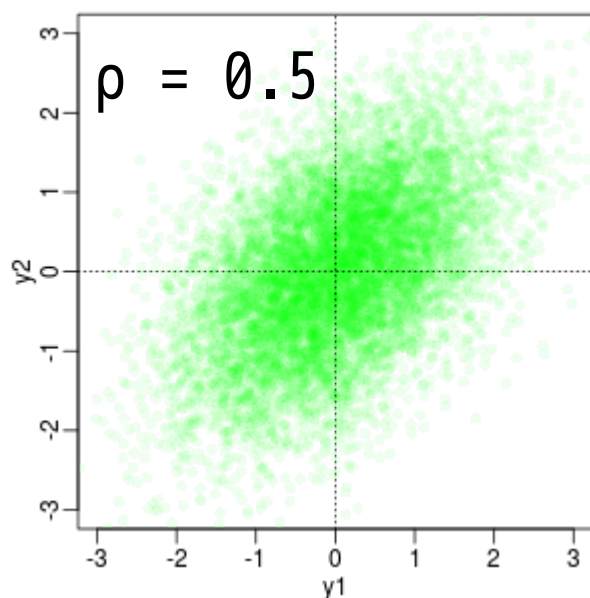
# 追加スライド

## 状態空間モデルの パラメータの 事後分布 ( $\rho = 0.0$ )



# 追加スライド

## 状態空間モデルの パラメータの 事後分布 ( $\rho = 0.5$ )



# 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

# 統計モデリング入門, ここまで…

データの性質・構造をよくみて統計モデルを作る

