

複数の時系列データの間を探る

— spurious correlation と spurious regression

粕谷英一

ランダムウォークから得られた時系列データ

独立

相関を見ると

ランダムウォークから得られた時系列データ

平均0の正規分布

時系列その1 $x_t = x_{t-1} + N(0, \sigma^2)$

ランダムウォークから得られた時系列データ

独立

ランダムウォークから得られた時系列データ

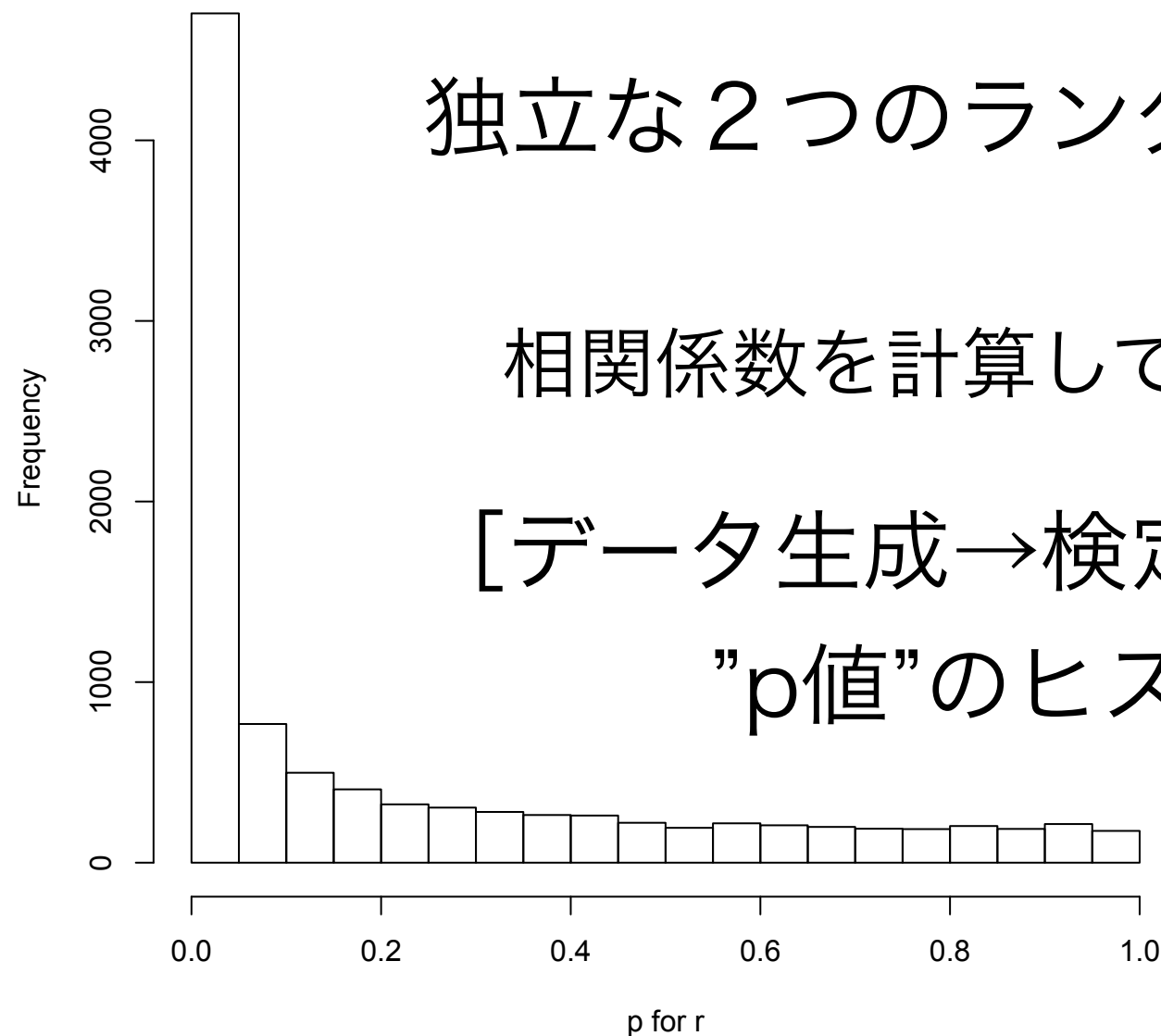
時系列その2 $y_t = y_{t-1} + N(0, \sigma^2)$

平均0の正規分布

相関係数 r の検定

n (時系列の長さ) = 20

誤って有意な相関と判定する率 (5%水準) = 0.4703



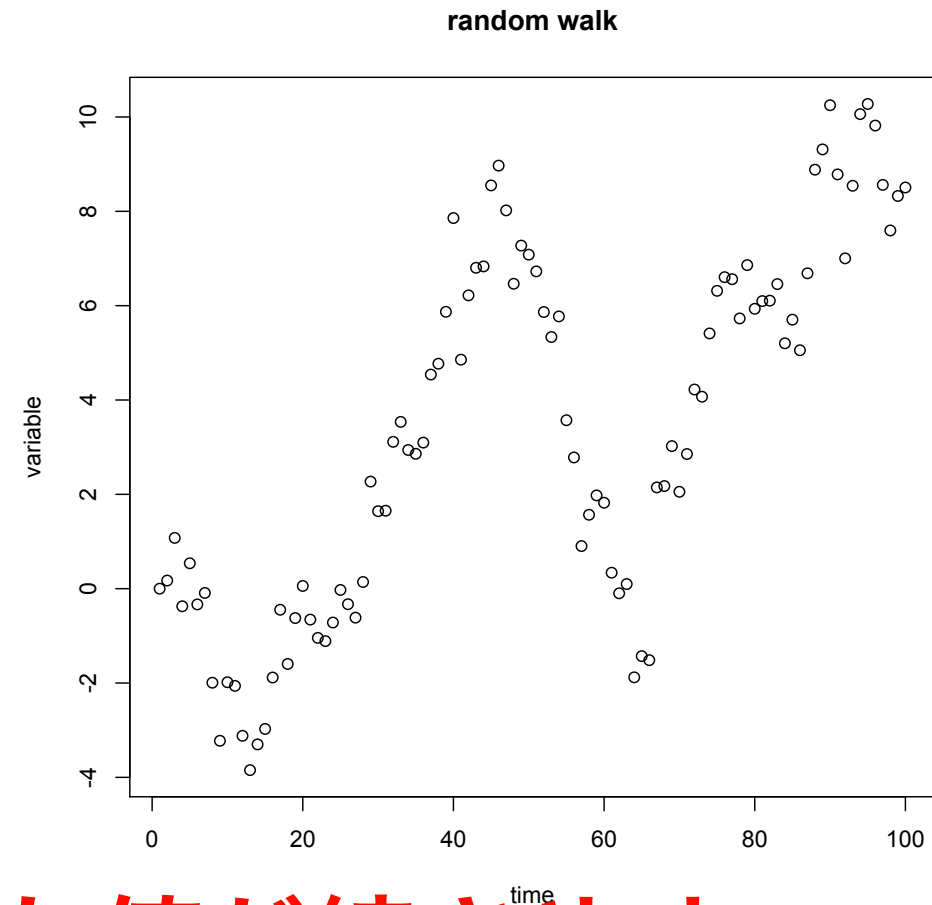
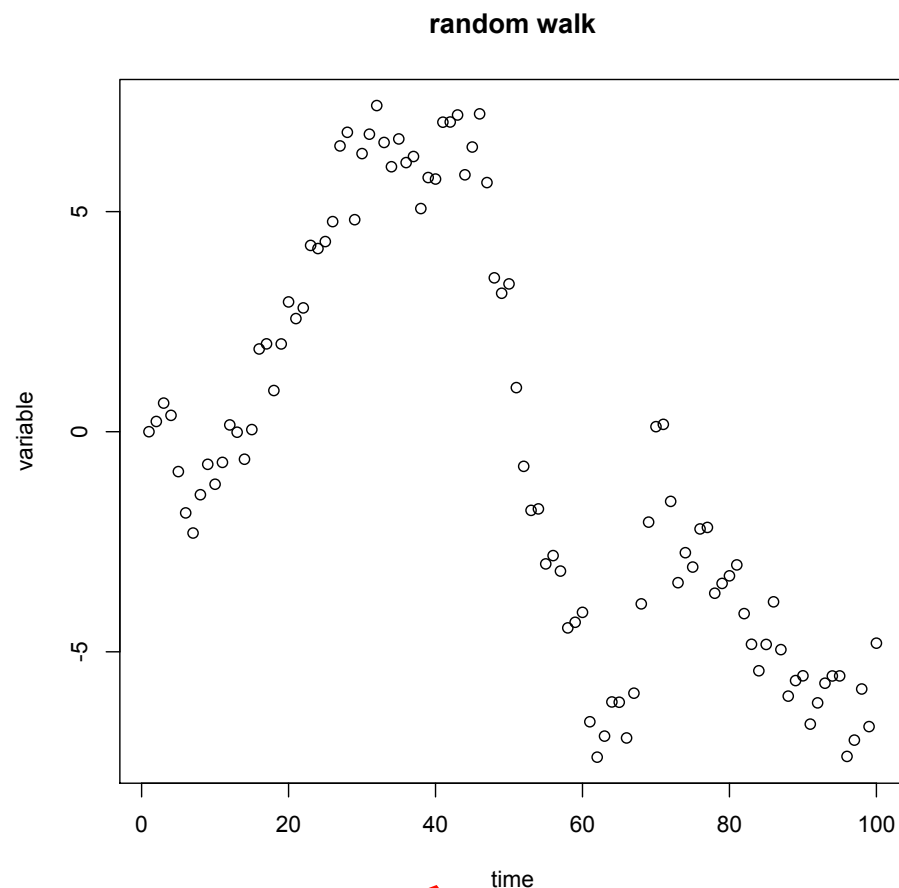
独立な2つのランダムウォーク

相関係数を計算して、検定した”p値”

[データ生成→検定] を繰り返して
”p値”のヒストグラム

pの値 (有意確率)

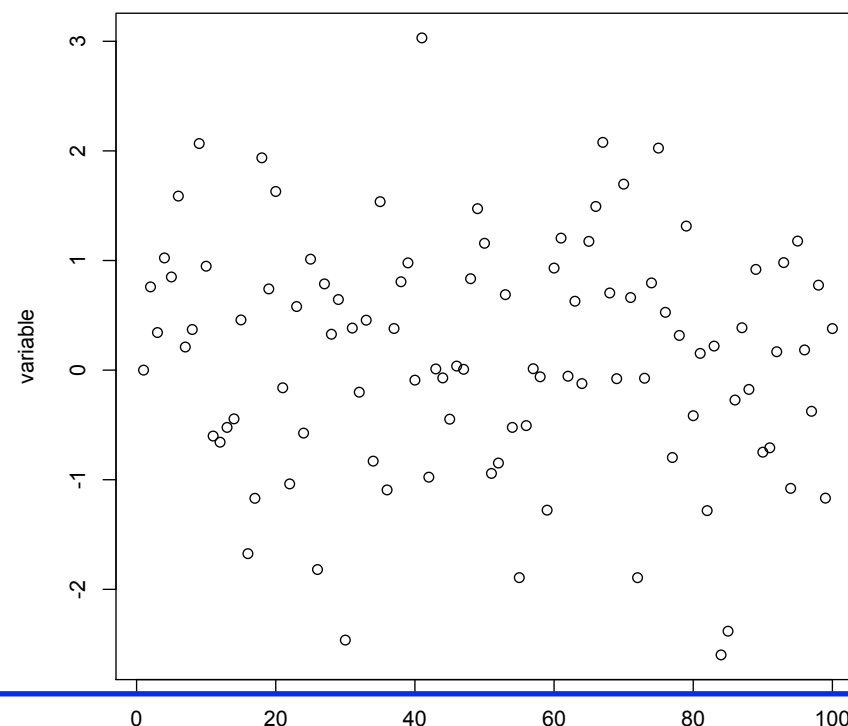
ランダム・ウォーク $x_t = x_{t-1} + N(0, \sigma^2)$



ランダム・ウォーク：似た値が続きやすい

i.i.d. normal

ある値が登場する
確率は、
前の値が何でも、
変わらない



正規分布 独立なデータ

$$x_t = N(0, \sigma^2)$$

経済時系列での研究

2つの独立なランダム・ウォーク

Granger & Newbold, 1974

Journal of Econometrics, 2: 111-120.

Phillips, 1986

Journal of Econometrics, 33: 311-340.

ある時点と次の時点の差（差分、階差）を使う

見せかけの相関

spurious correlation

見せかけの回帰

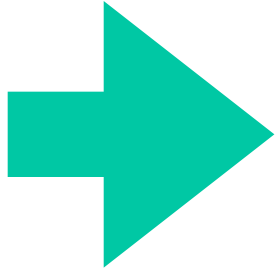
spurious regression

←より広い意味で、

変な相関全般を指すこと

もある

目次



- ・ 2つのランダムウォークの相関：再訪
- ・ 使う方法の説明

順位相関ではだめですか

ランダムウォークの動きの大きさ（分散）がちがうと
正規分布でないと

ランダムウォークと周期的変数だと

- ・ サンプルサイズが大きくなると
- ・ 経済時系列でわかっていること
- ・ 回帰の場合

データ生成
互いに独立

時系列 x

時系列 y

相関を検定

相関係数 r

とくに言わなければ、ランダム・ウォーク

10000回行う

サンプルサイズ (n) 時系列の長さ

5%水準で有意と判定される回数を数える

互いに独立なプロセスで生成されたデータ

→有意な相関という判定：誤り

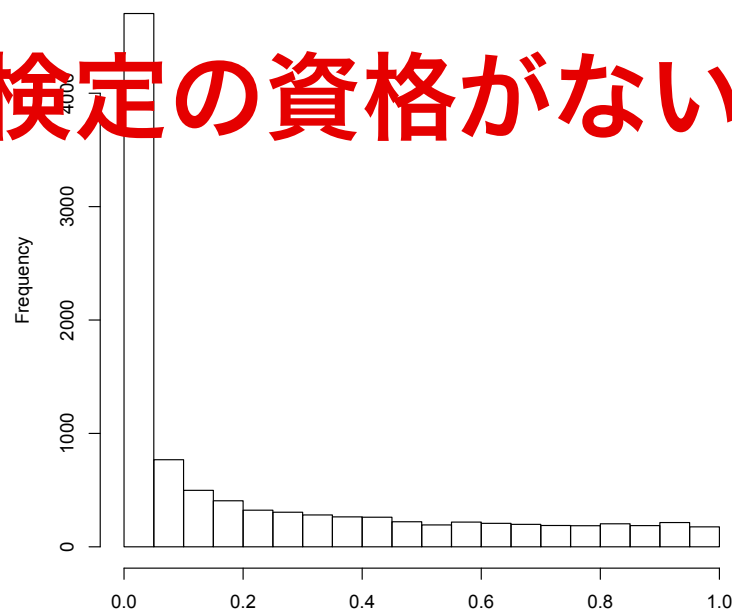
誤って有意と判定してしまう：第1種の誤り（第1種の過誤）

”p値”（有意確率）のヒストグラム

甘すぎる
妥当でない

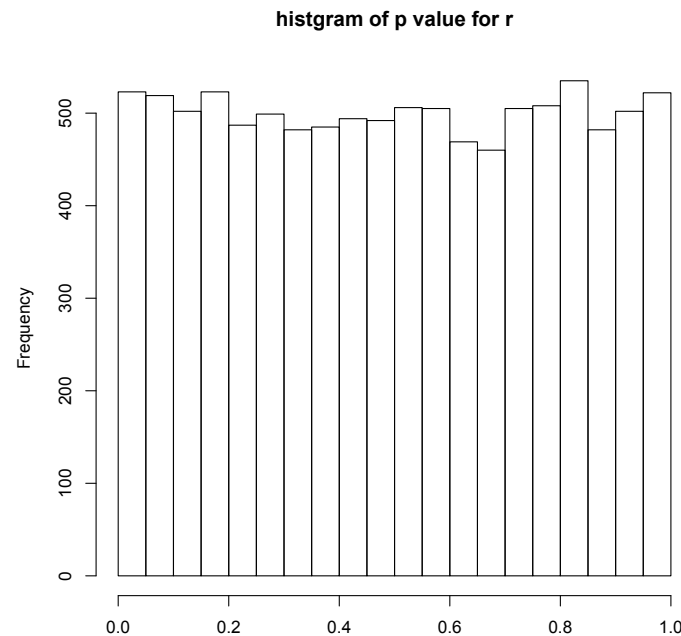
n=20 type1error(5%)=0.4703 no trend

”検定の資格がない”

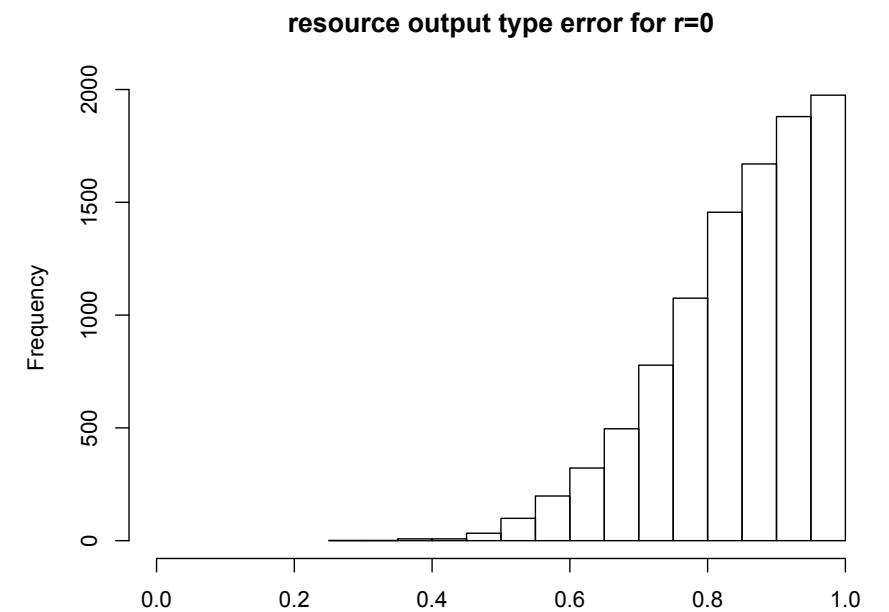


うまく行っている

有意水準として宣言した通りの第1種の誤りの率



きびしすぎる
（鈍感）



”p値”（有意確率）



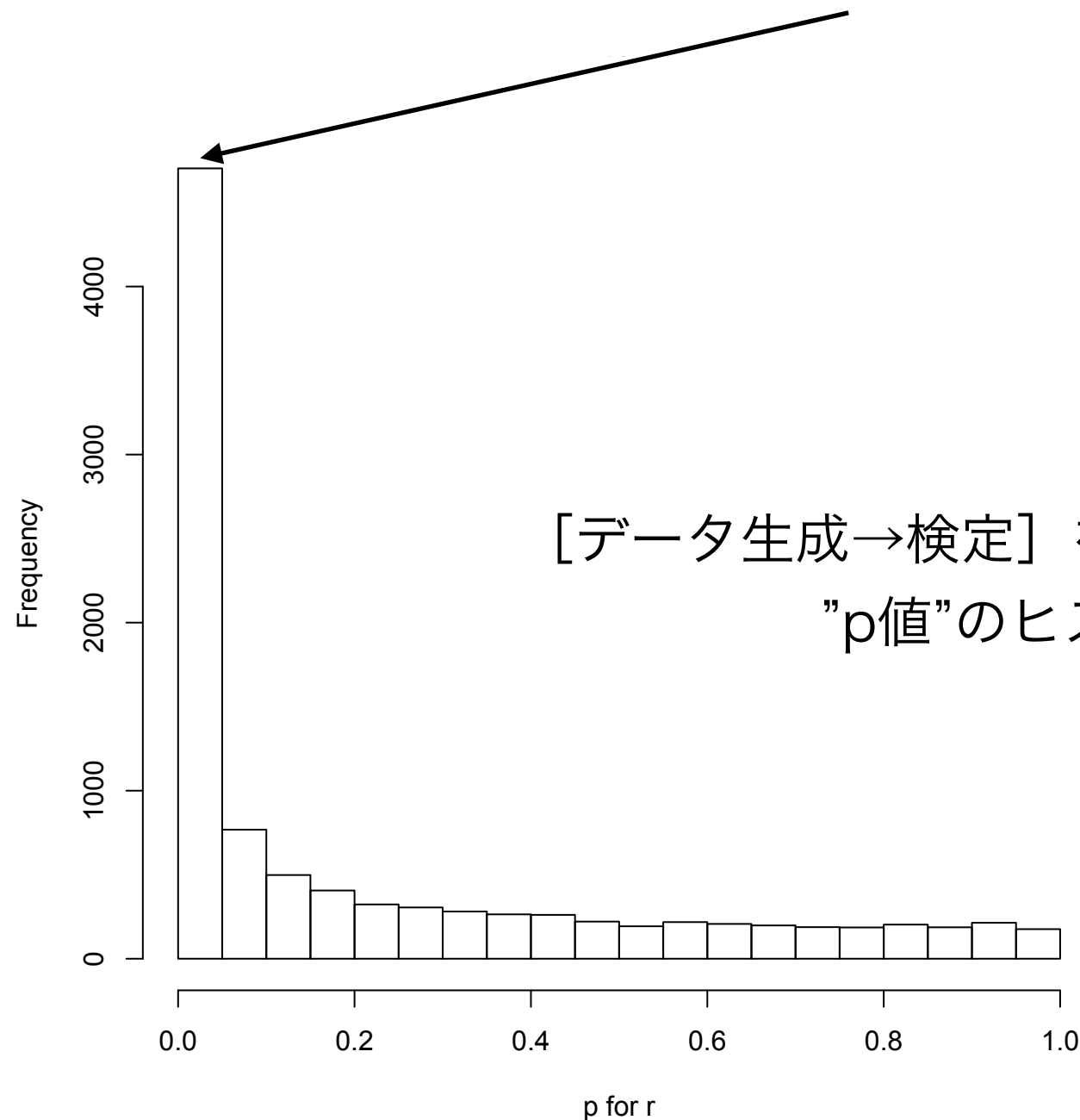
2つの無相関な正規分布からの独立なデータ
→相関係数の検定

独立な2つのランダムウォーク

相関係数 r の検定

誤って有意な相関と判定する率 (5%水準) = 0.4703

n (時系列の長さ) = 20



[データ生成→検定] を10000回繰り返して
"p値"のヒストグラム

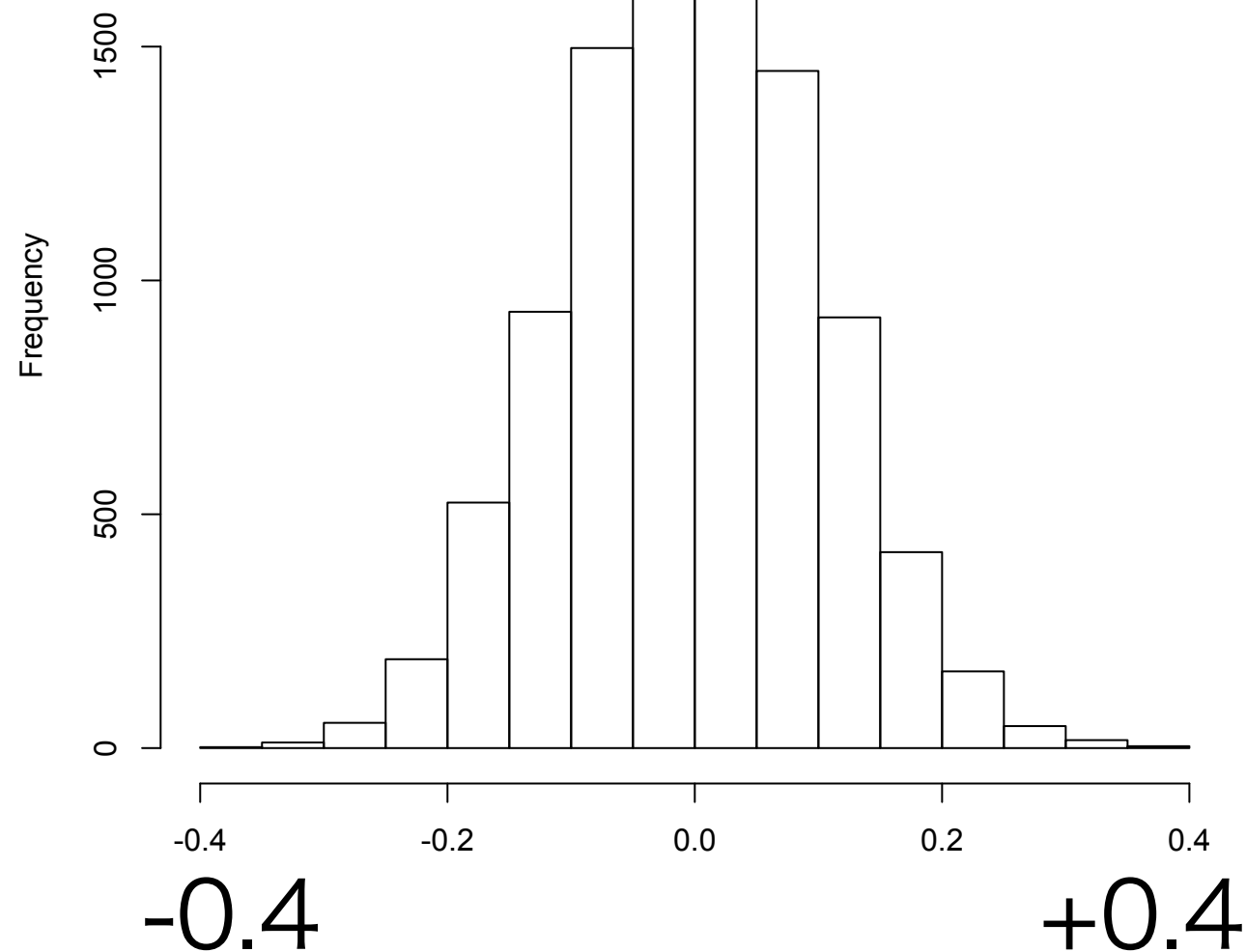
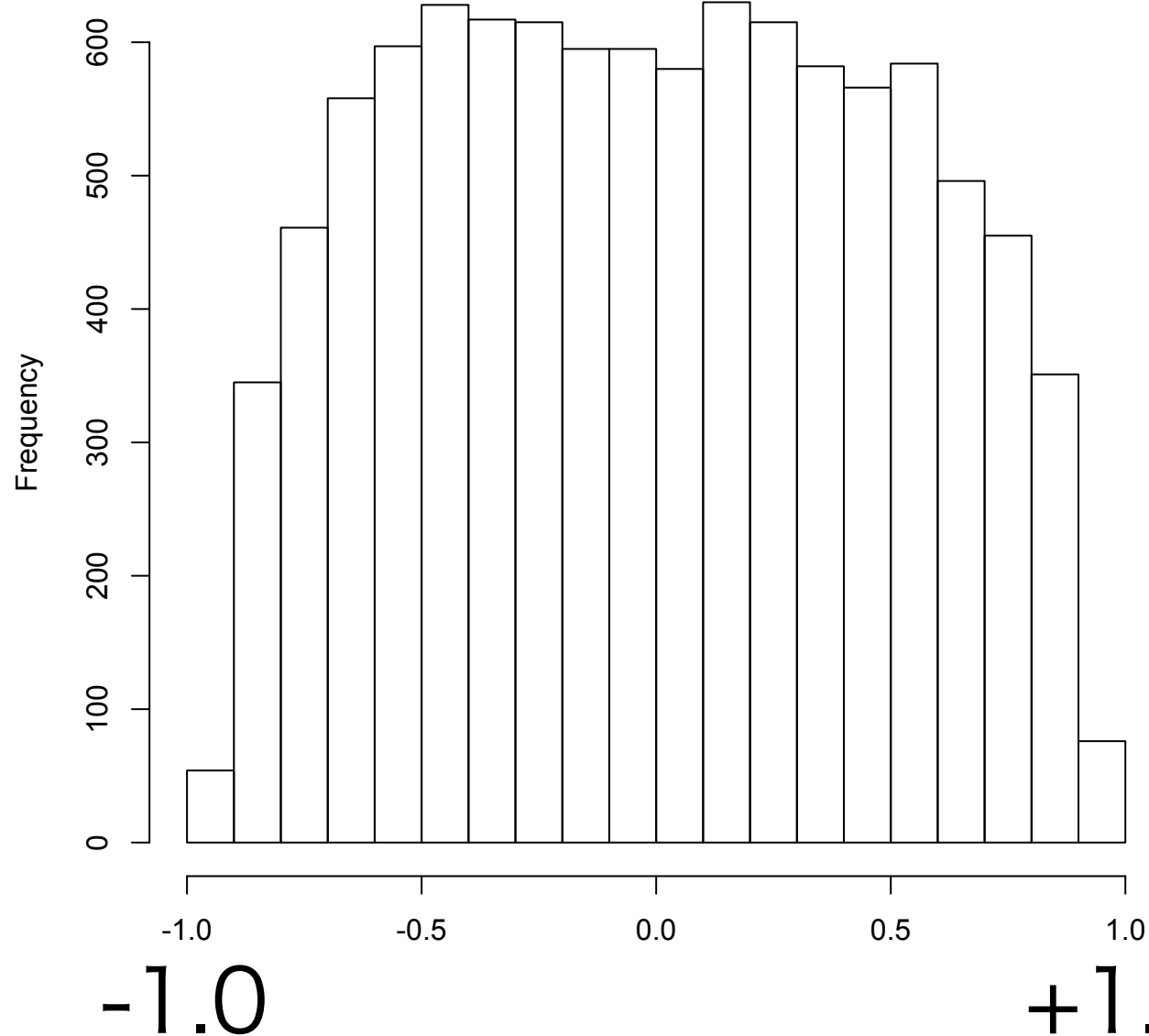
**甘すぎる
妥当でない**

pの値 (有意確率)

相関係数のヒストグラム

ランダム・ウォーク

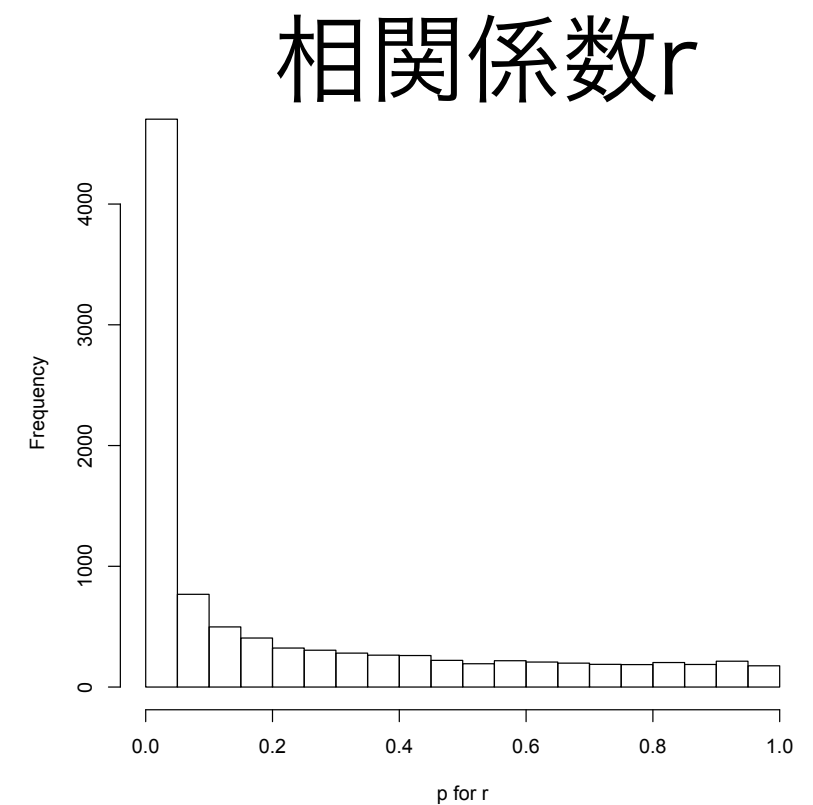
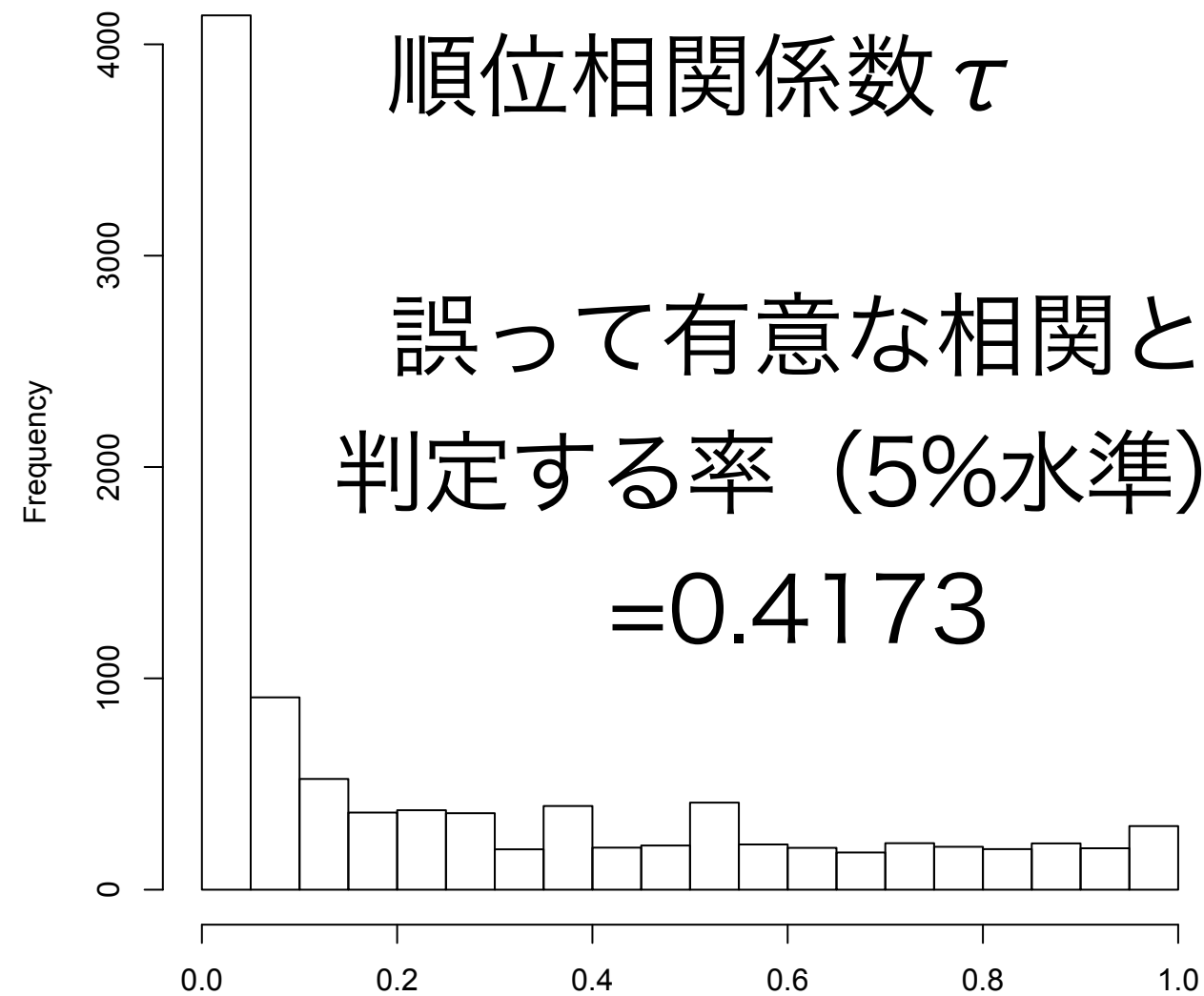
正規分布からの
独立なデータ



相関係数 r

順位相関係数でもほとんど変わらない

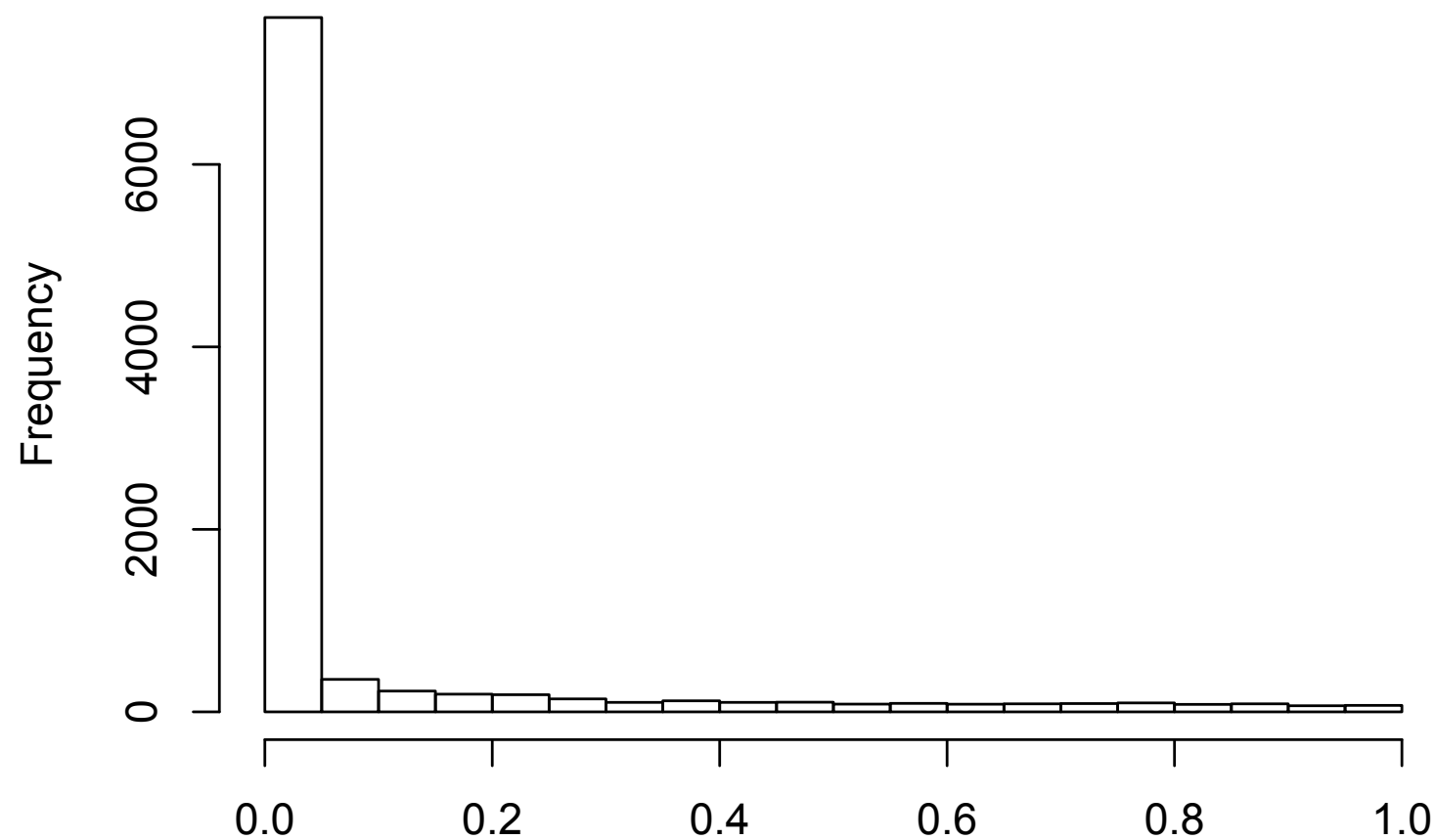
n（時系列の長さ）=20



”p値”（有意確率）

分散が異なっても、起こる

ランダムウォークの正規分布の分散 y は、 x の25倍



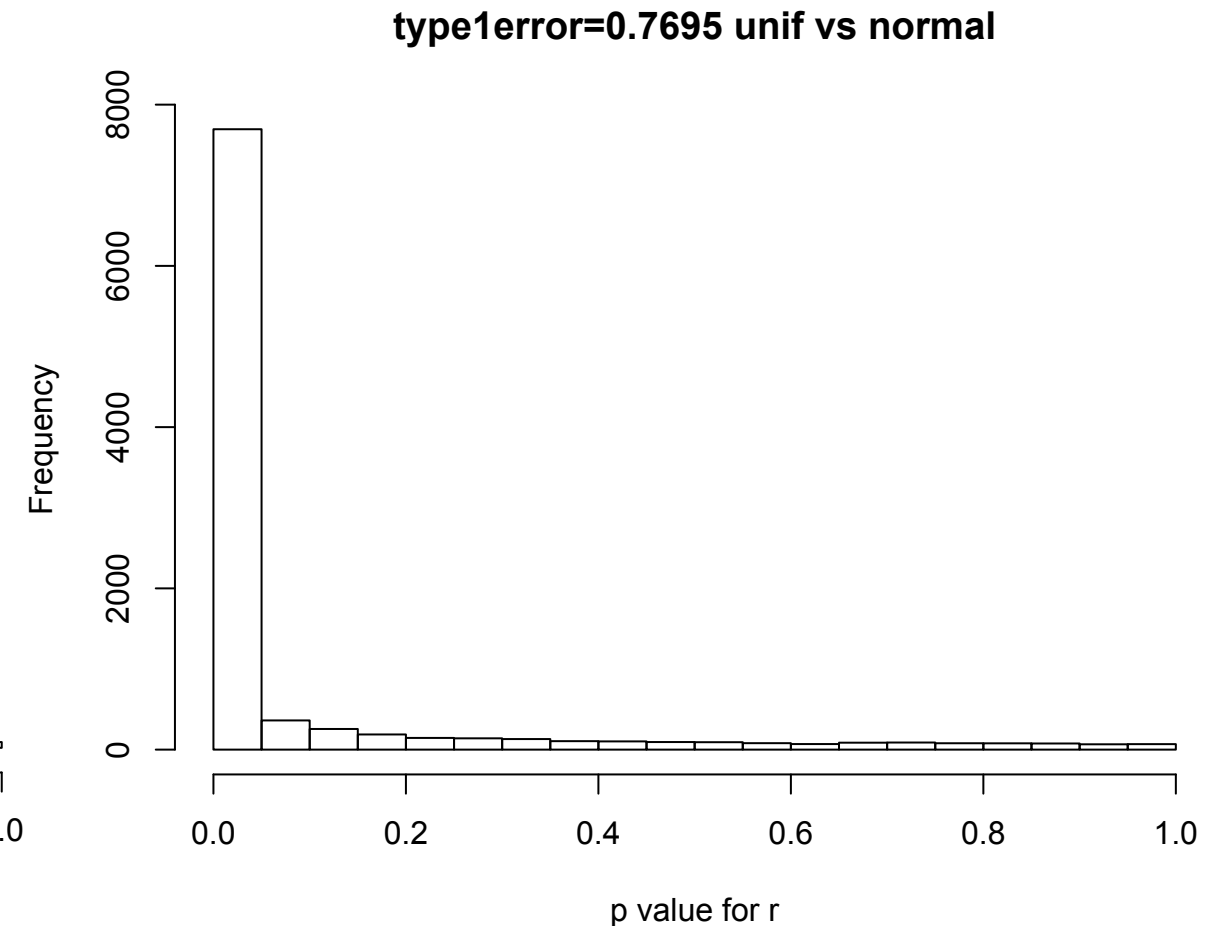
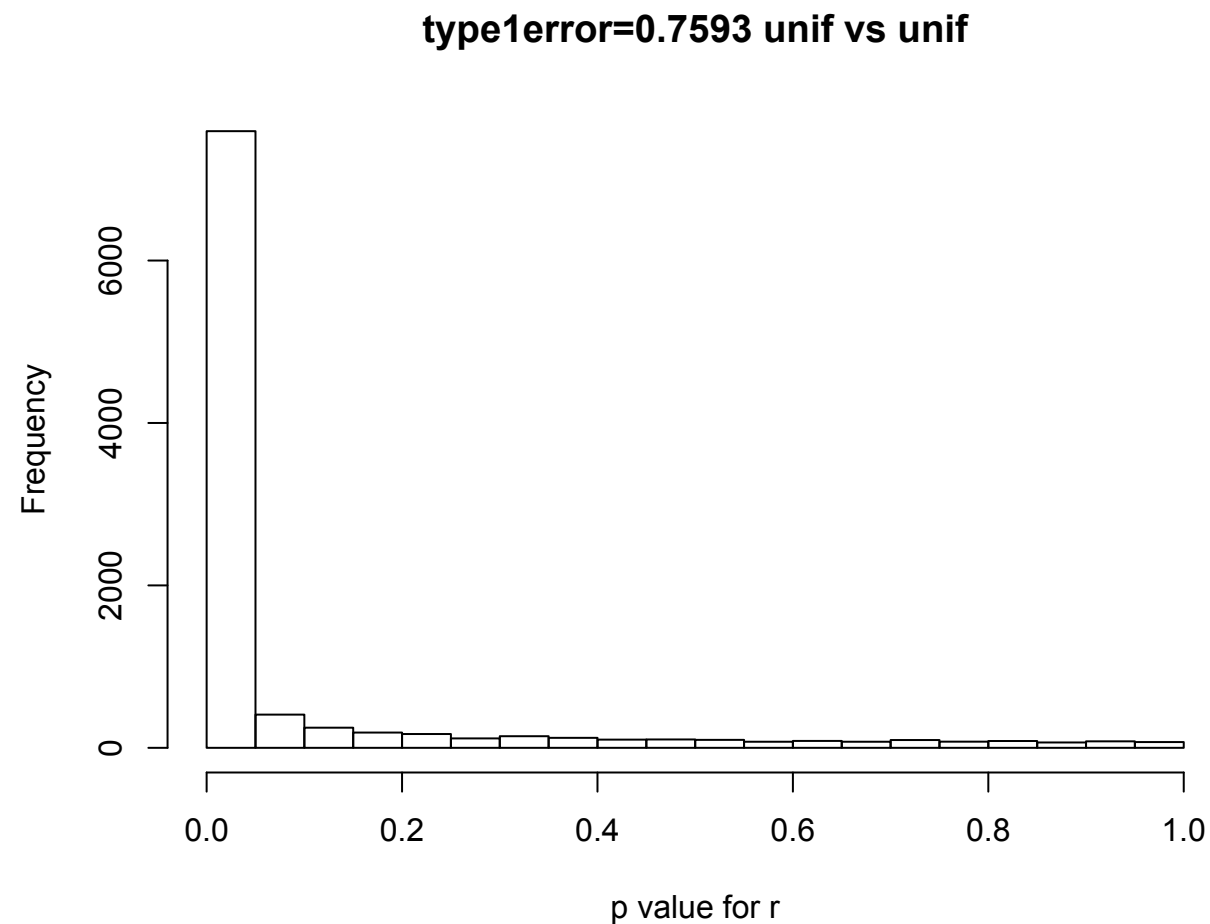
”p値” (有意確率)

正規分布でなくとも、起こる：一様分布の場合

n（時系列の長さ）=100

両方（x,y）とも一様分布

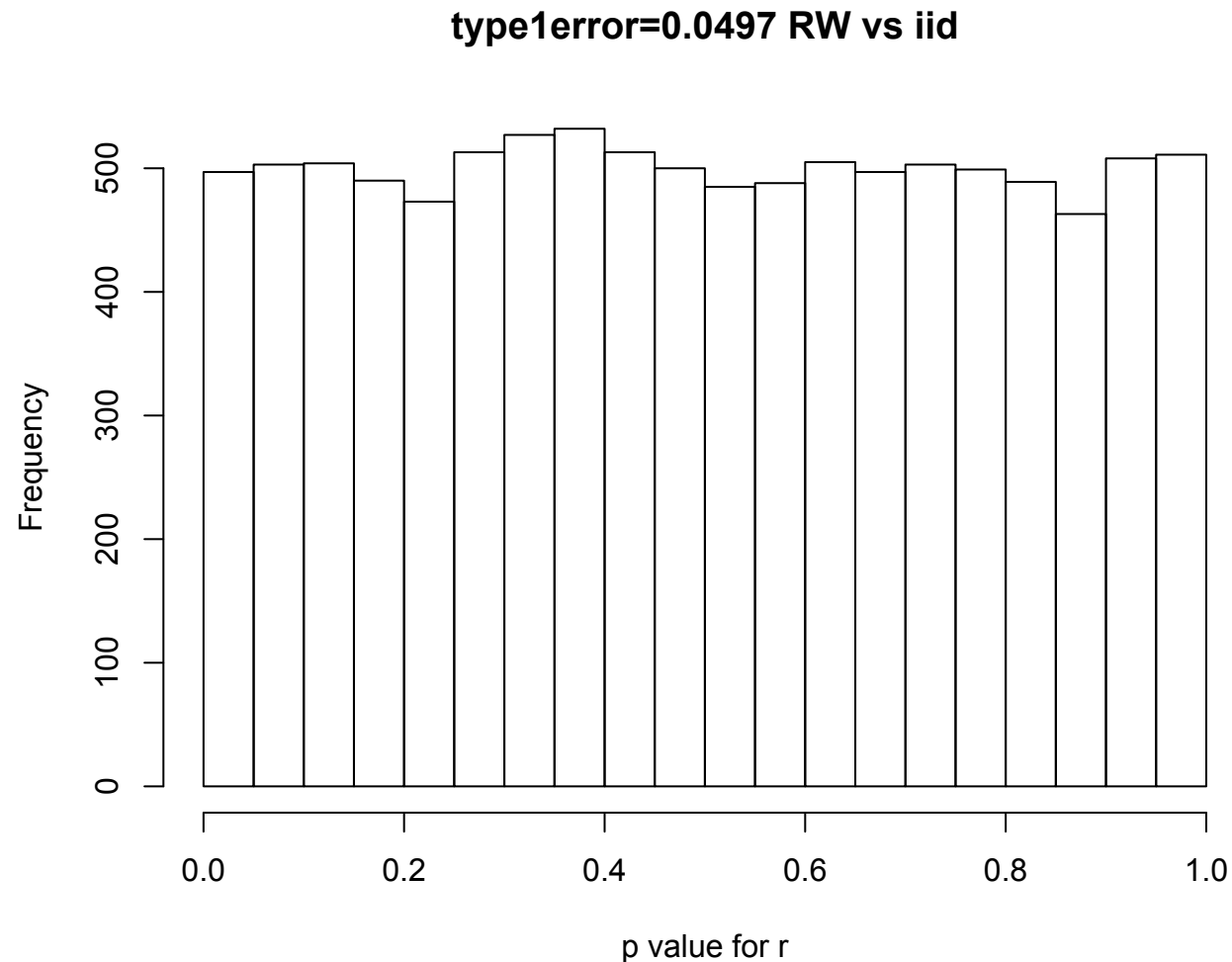
正規分布と一様分布



pの値（有意確率）

片方がランダム・ウォーク 片方が正規分布からの独立なデータ

n（時系列の長さ）=100



pの値（有意確率）

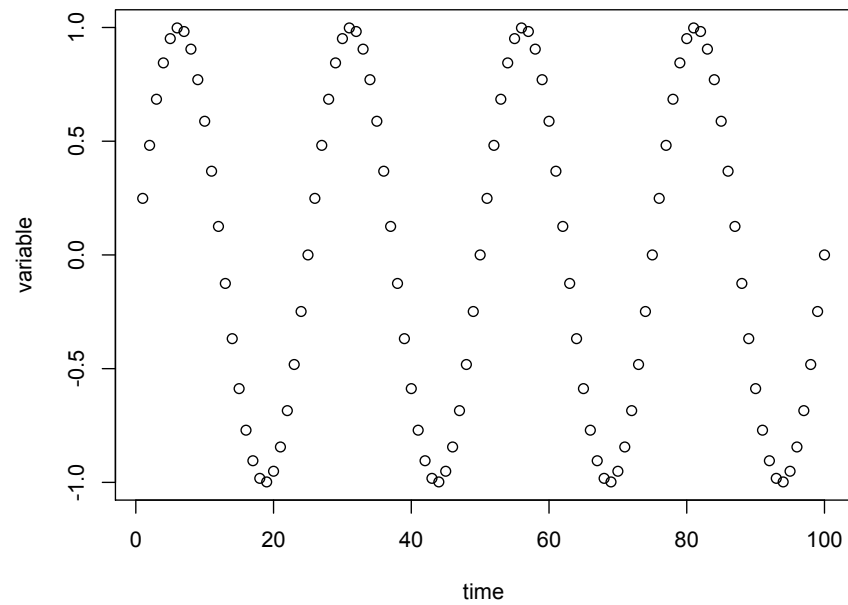
このときは、問題なし

片方が周期的変数、もう片方がランダムウォーク

ここではサインカーブを使った

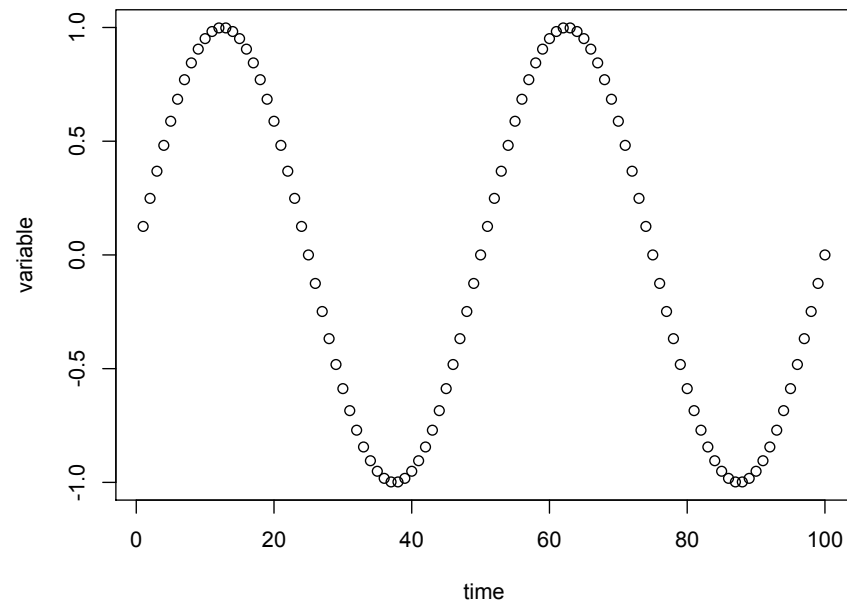
4周期

sine curve 4cycles



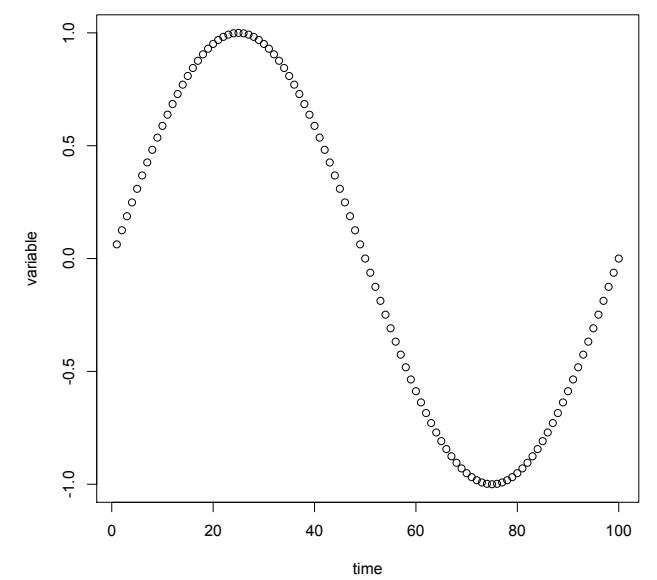
2周期

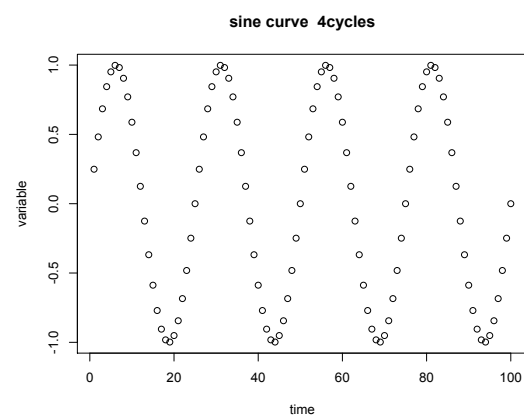
sine curve 2cycles



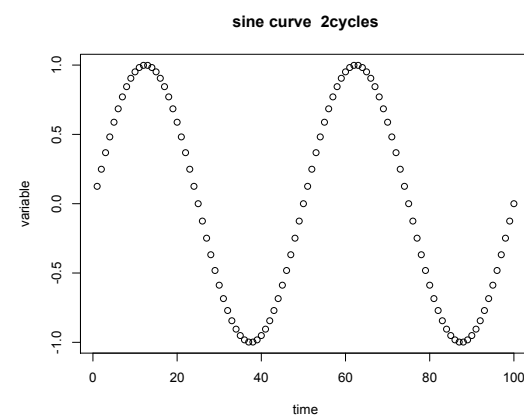
1周期

sine curve 1 cycle

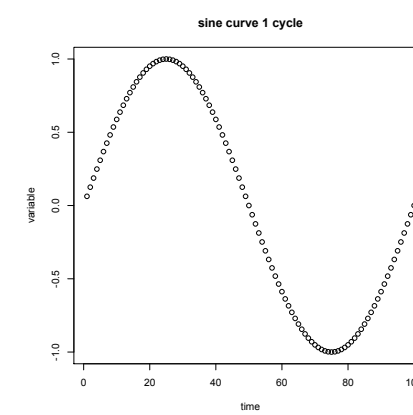




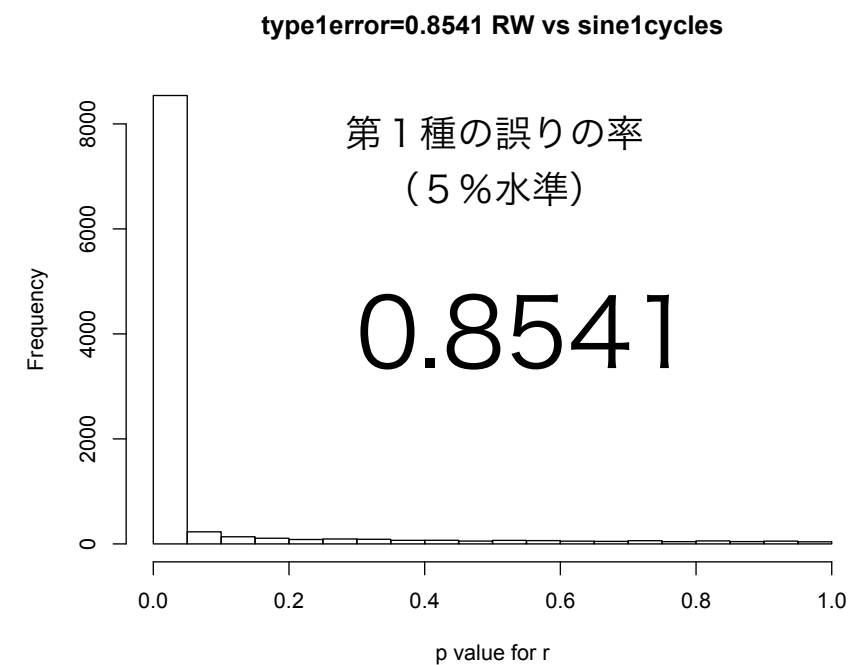
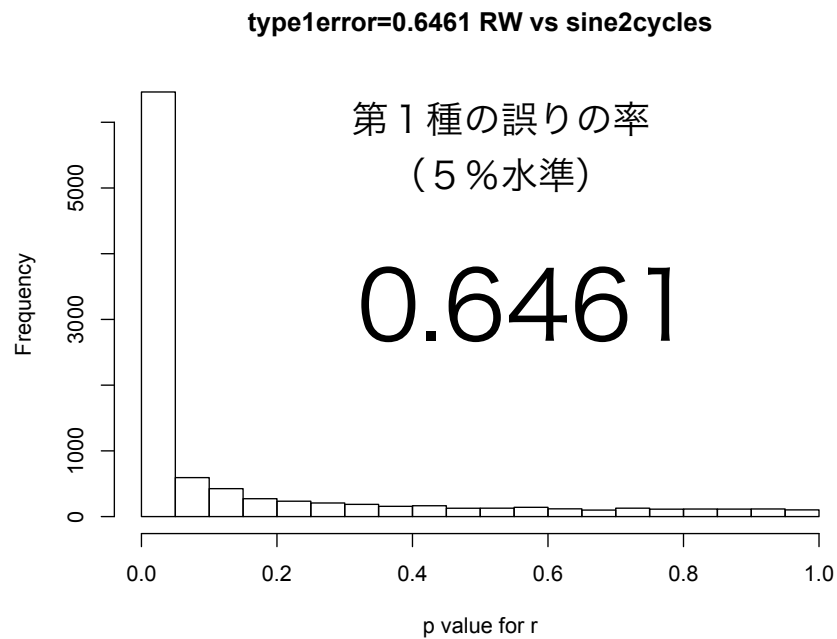
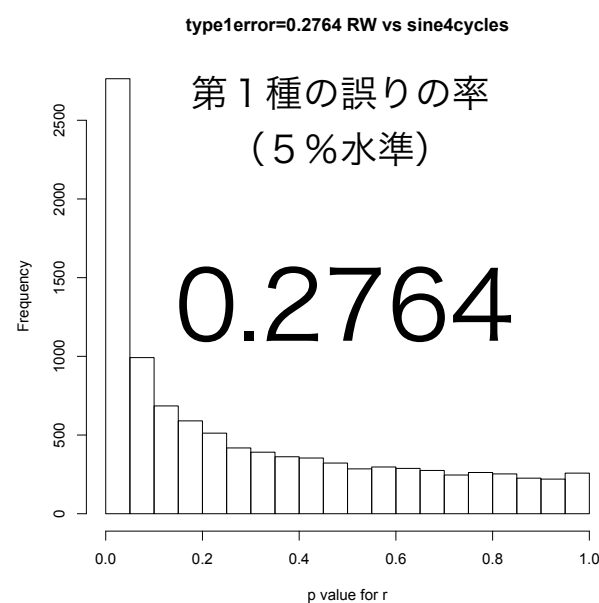
4周期



2周期



1周期

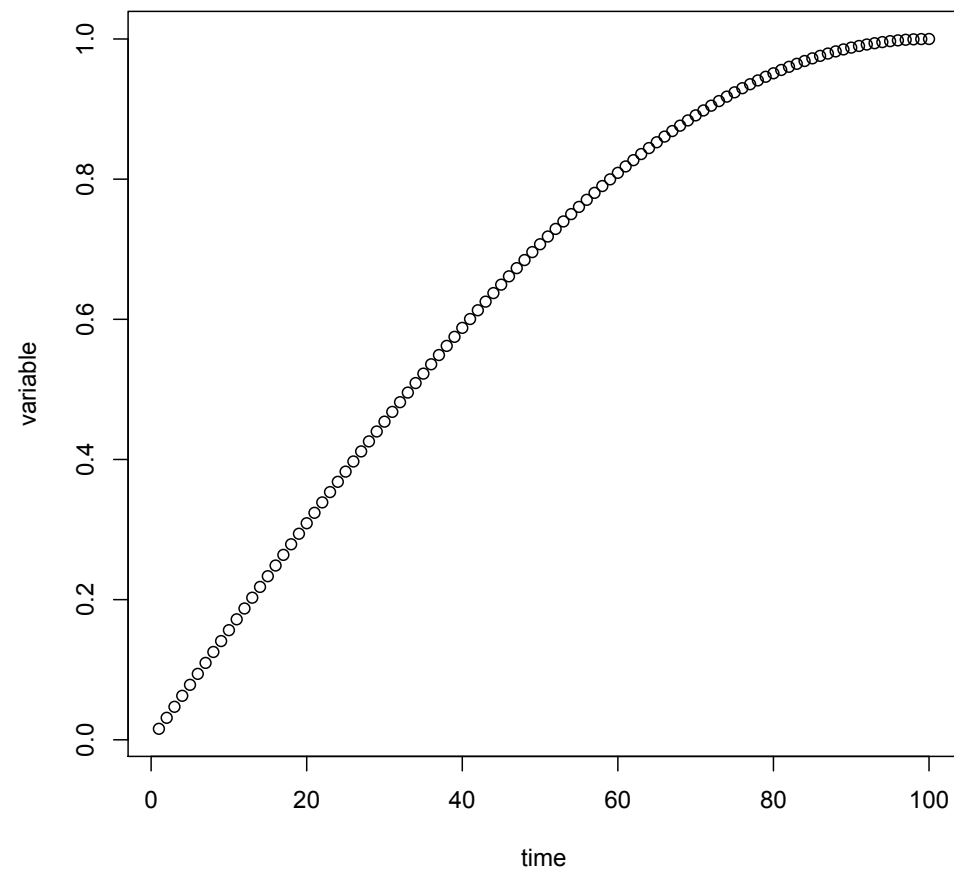


”p値”（有意確率）

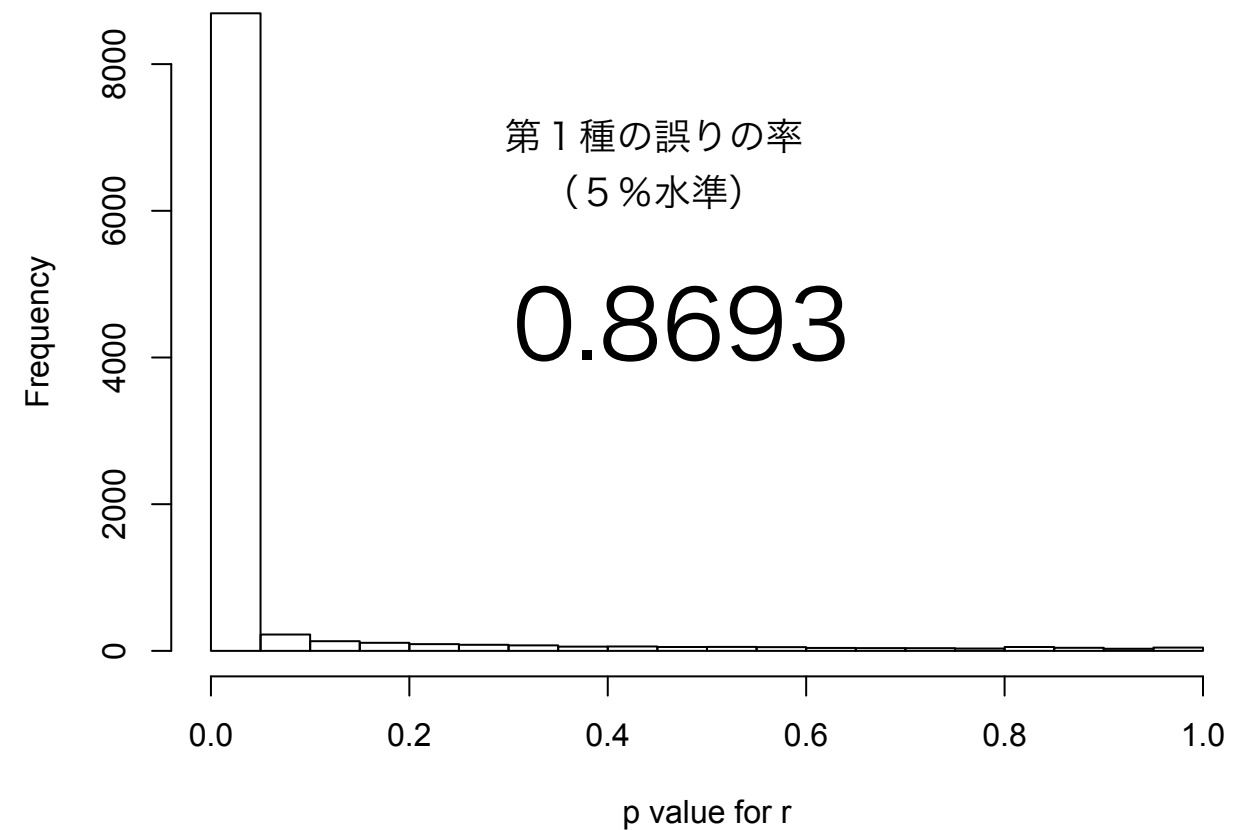
片方がサインカーブの1/4周期分 もう片方がランダムウォーク

1/4周期

sine curve quarter cycle

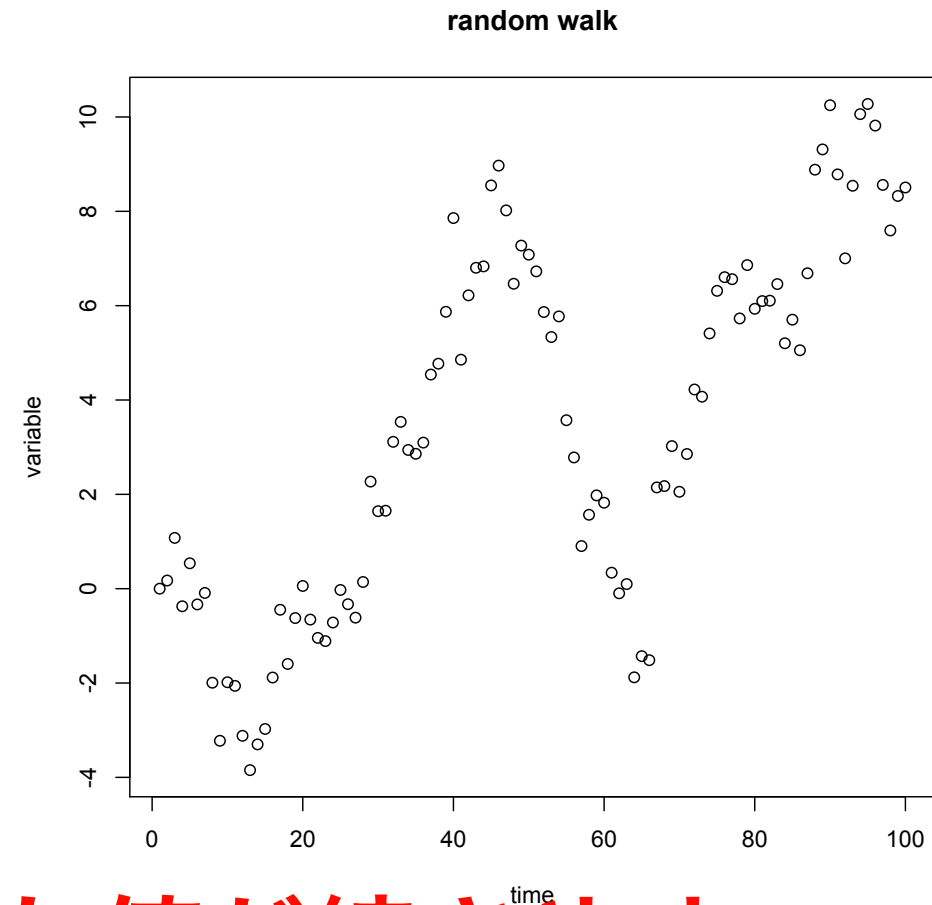
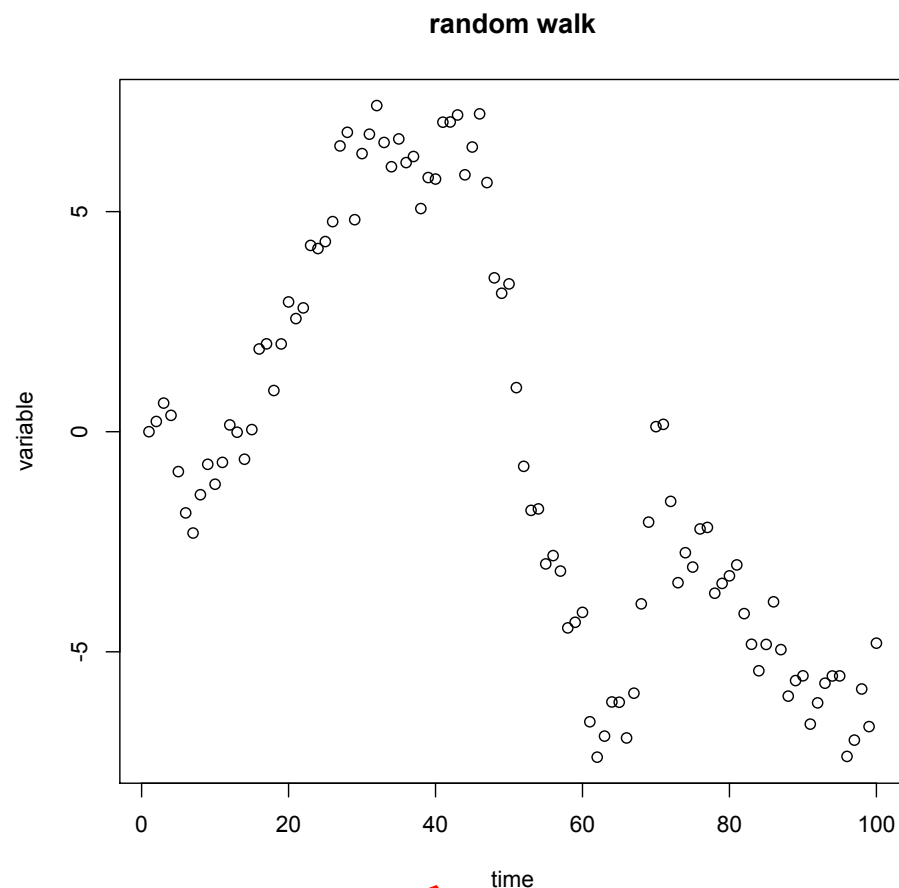


type1error=0.8693 RW vs sine1/4cycle



”p値”（有意確率）

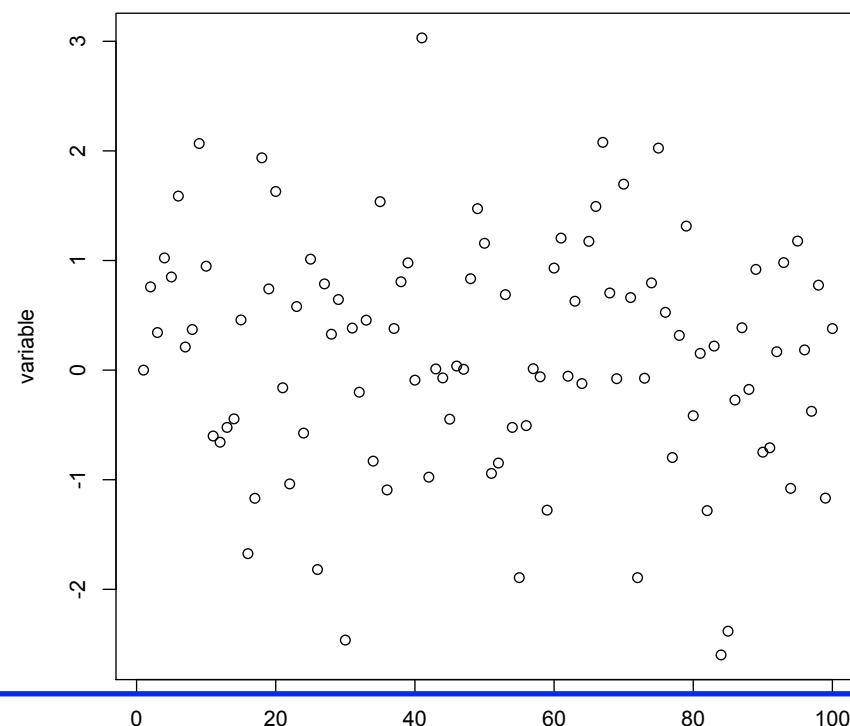
ランダム・ウォーク $x_t = x_{t-1} + N(0, \sigma^2)$



ランダム・ウォーク：似た値が続きやすい

i.i.d. normal

ある値が登場する
確率は、
前の値が何でも、
変わらない



正規分布 独立なデータ

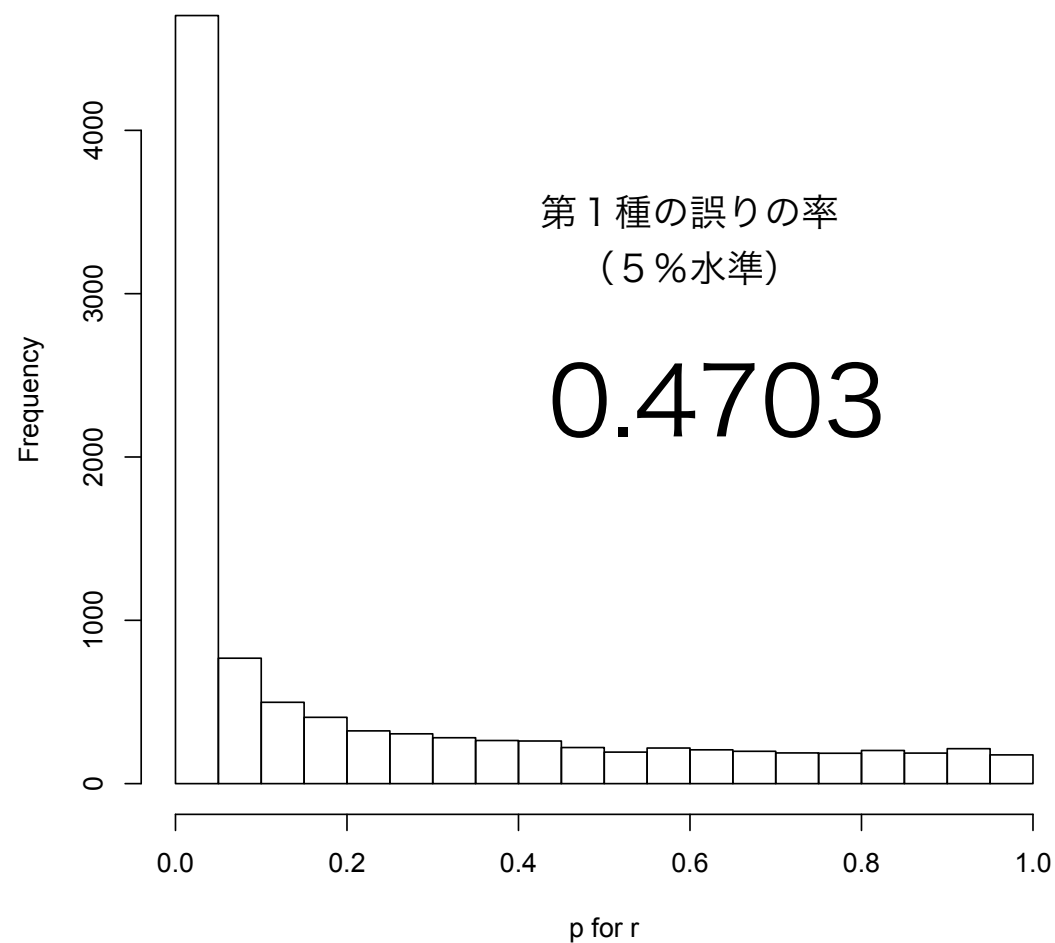
$$x_t = N(0, \sigma^2)$$

似た値が続きやすい

サンプルサイズの影響 時系列の長さ

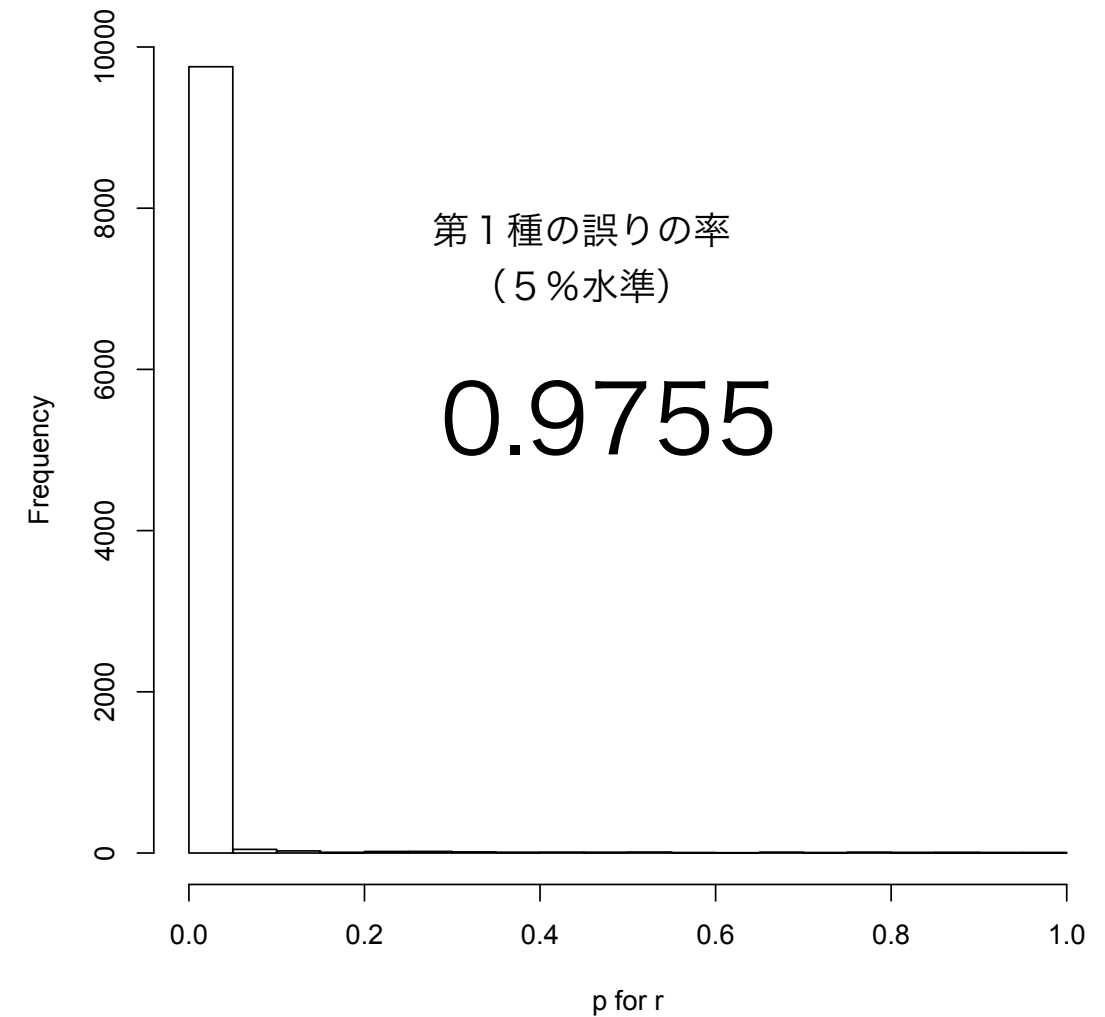
$n=20$

n=20 type1error(5%)=0.4703 no trend



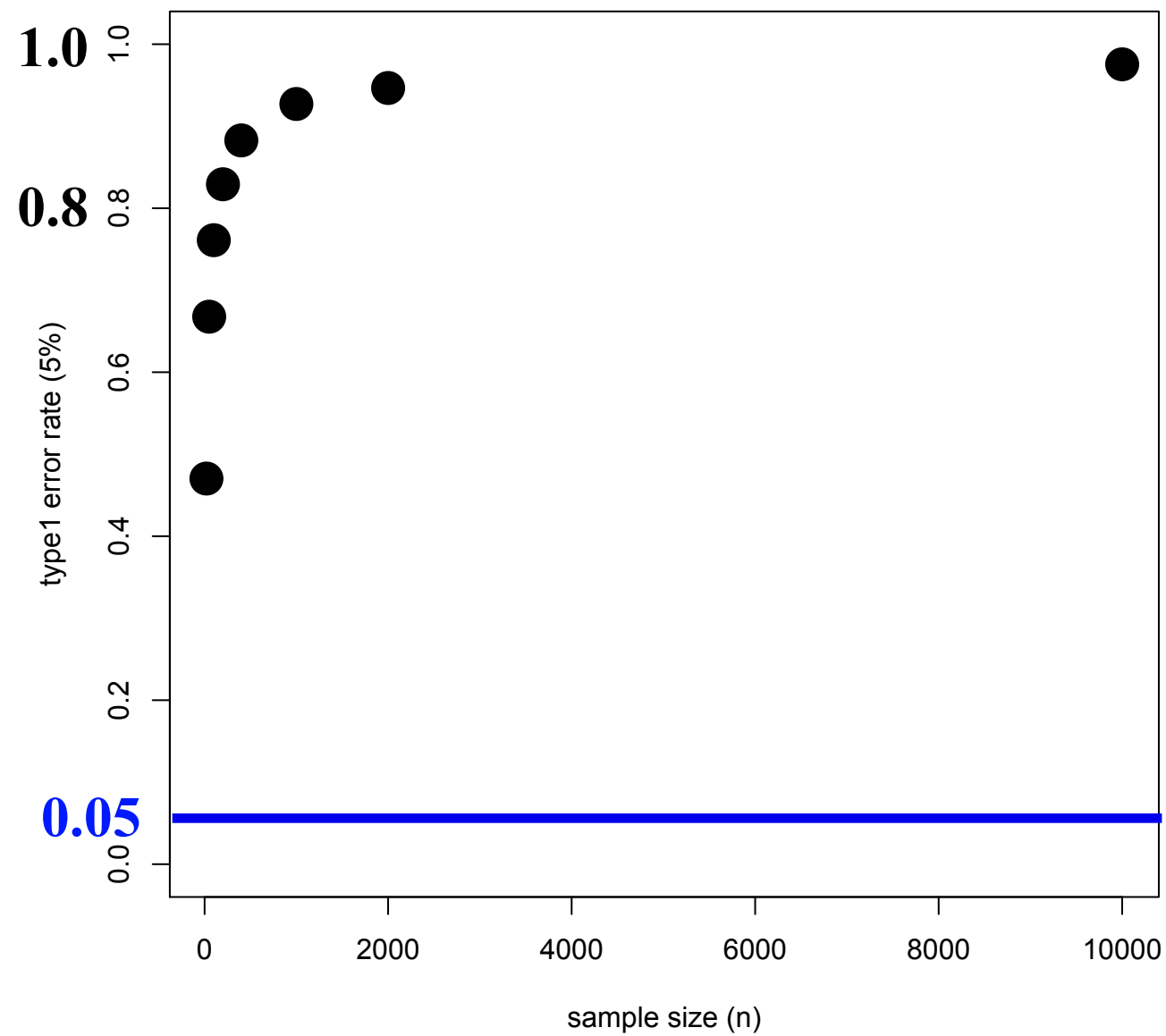
$n=10000$

n=10000 type1error(5%)=0.9755 no trend



”p値” (有意確率)

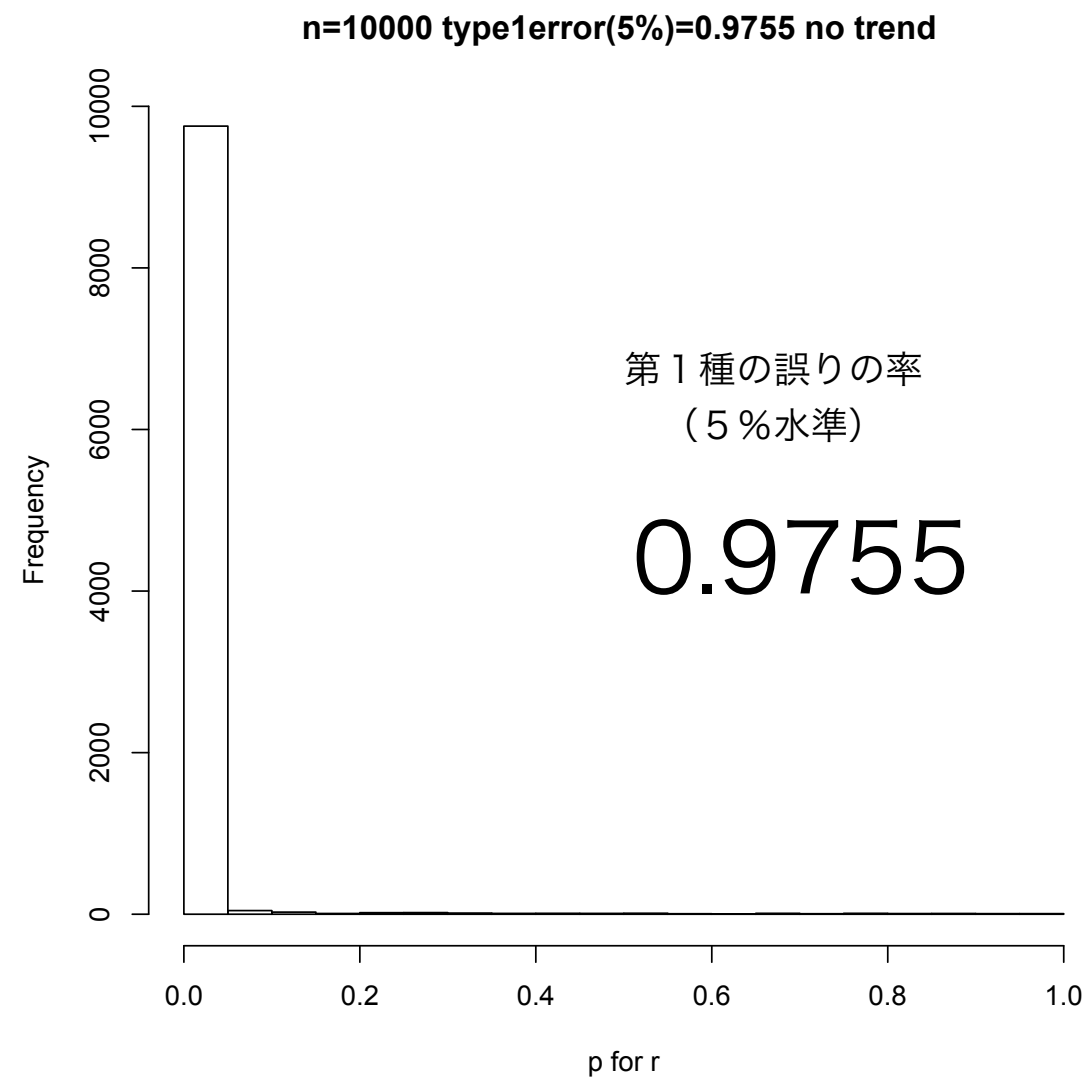
5%水準で検
定した時の
第1種の誤り
の率



サンプルサイズ(n)

時系列の長さ

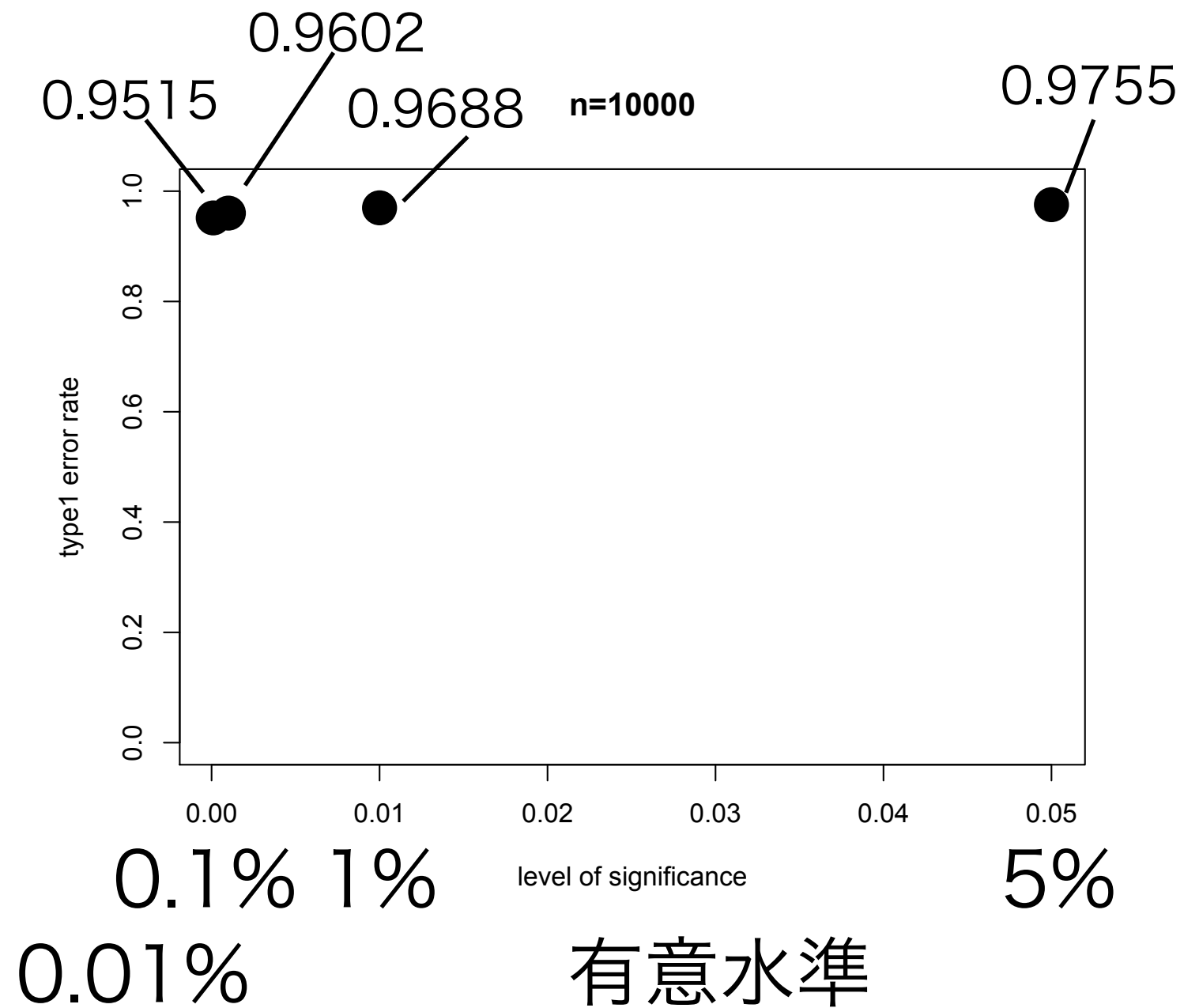
$n=10000$



”p値” (有意確率)

$n=10000$

第1種の
誤りの率



サンプル・サイズが大きいと、
(時系列の長さ)
さらに悪化する

経済時系列での研究

見せかけの相関

spurious correlation

見せかけの回帰

spurious regression

Granger & Newbold, 1974

Journal of Econometrics, 2: 111-120.

Phillips, 1986

Journal of Econometrics, 33: 311-340.

単位根過程

定常過程

期待値が一定

時間間隔が同じなら、

自己相関係数の期待値が

どこでも一定

非定常過程

$$x_t = \theta_x x_{t-1} + N(0, \sigma^2)$$

定常過程

θ_x の絶対値 < 1

非定常過程

$\theta_x = 1$ 単位根過程

ランダムウォーク

定常過程

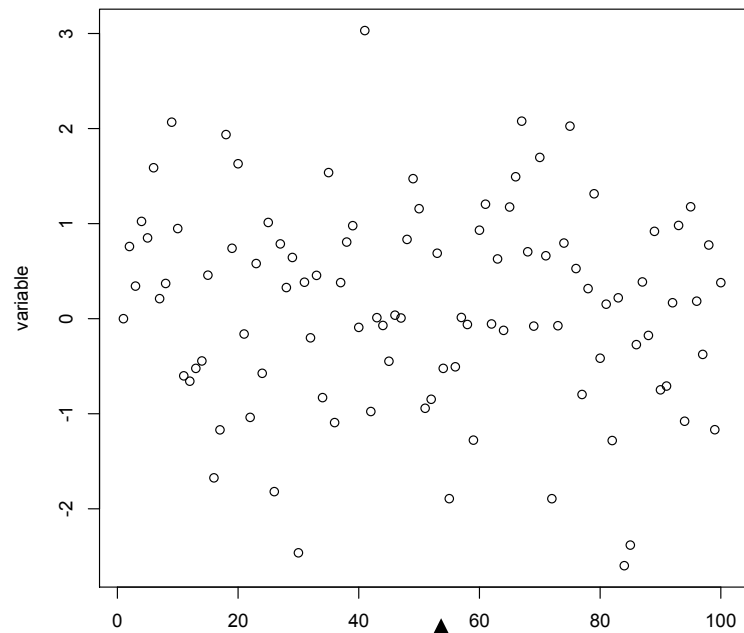
例

感じがちがう

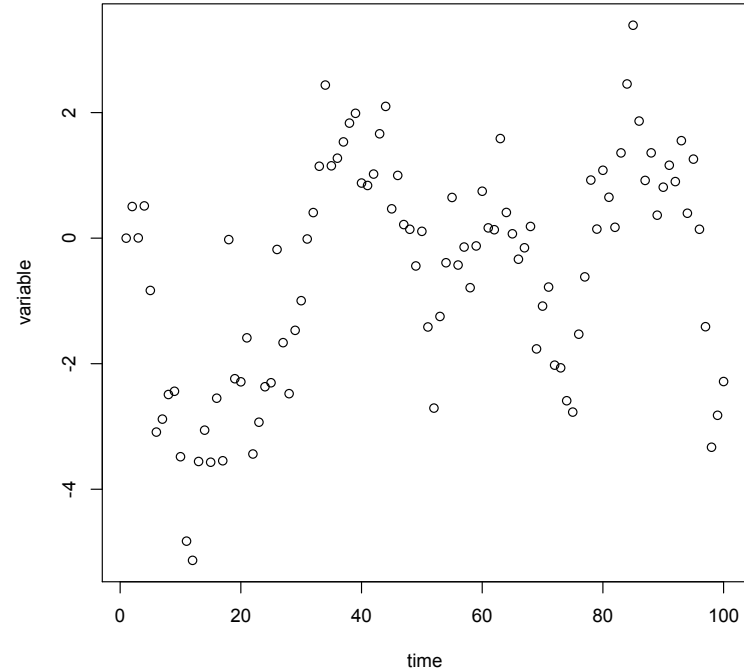
非定常過程

單位根過程

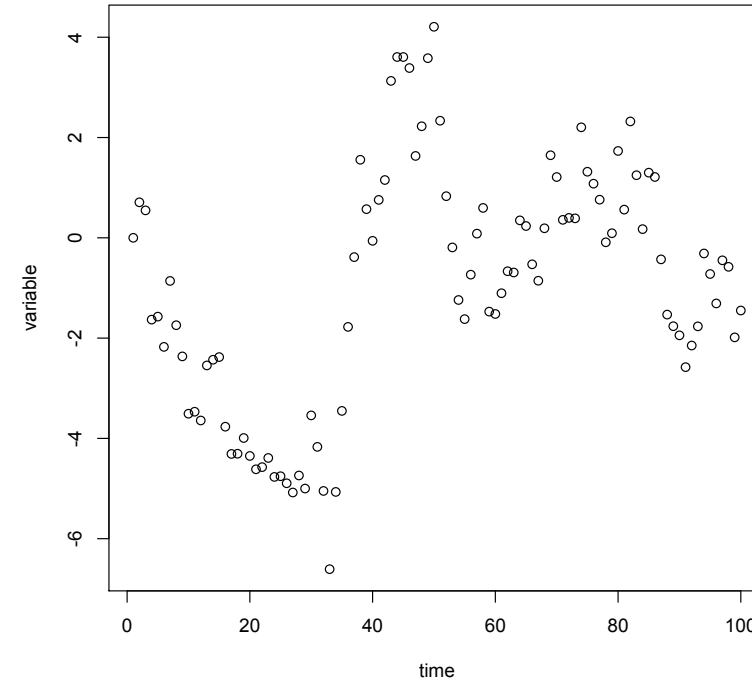
i.i.d. normal $\theta x = 0.00$



coef=0.95 $\theta x = 0.95$



coef=0.98 $\theta x = 0.98$

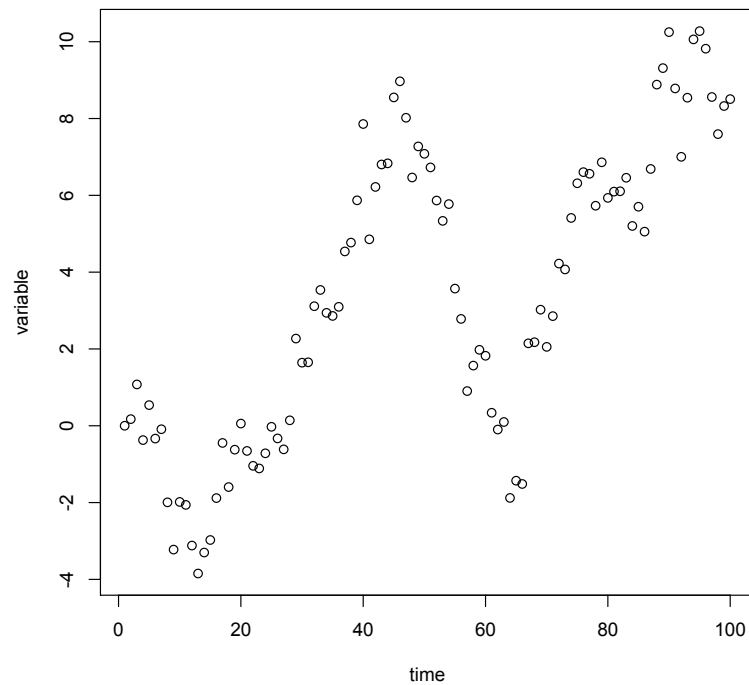


定常

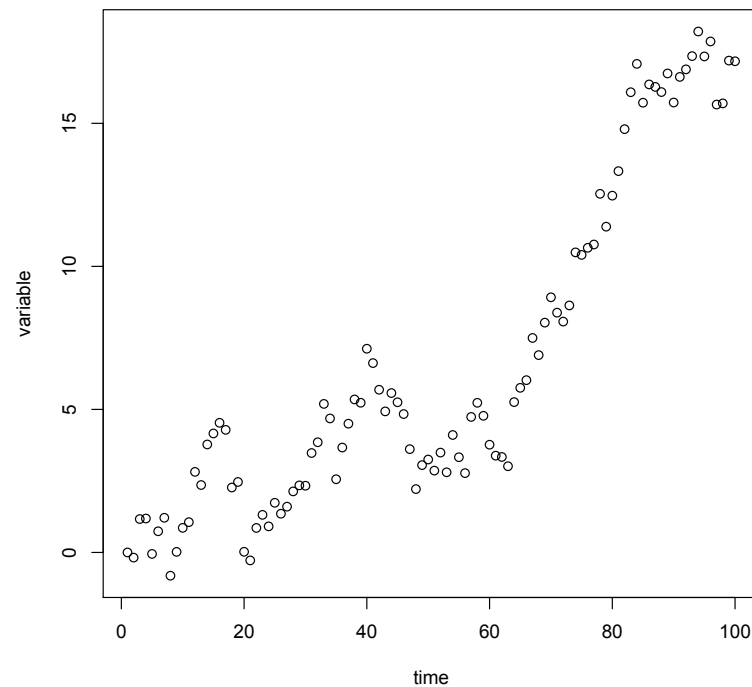
正規分布からの
独立なデータ

時間

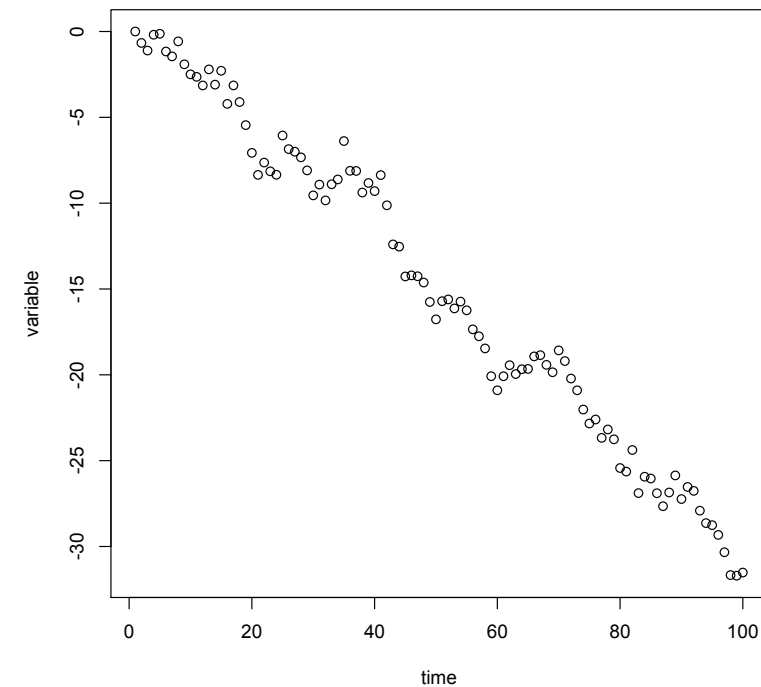
ランダムウォーク $\theta x = 1.00$



coef=1.01 $\theta x = 1.01$



coef=1.02 $\theta x = 1.02$

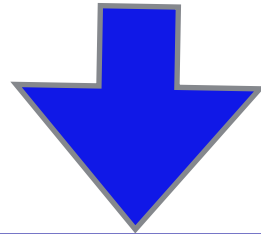


非定常

時間

2つの単位根過程

たとえば、ランダム・ウォーク



見せかけの相関

spurious correlation

強い相関が出やすい

定常過程

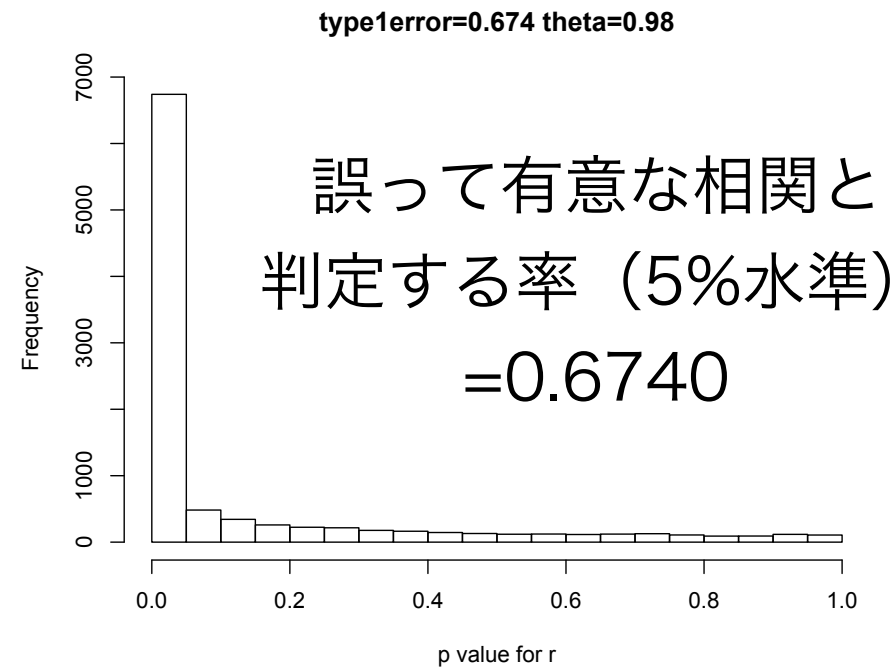
Granger et al (2001) Applied Economics, 33: 899-904.

$$x_t = \theta_x x_{t-1} + N(0, \sigma^2)$$

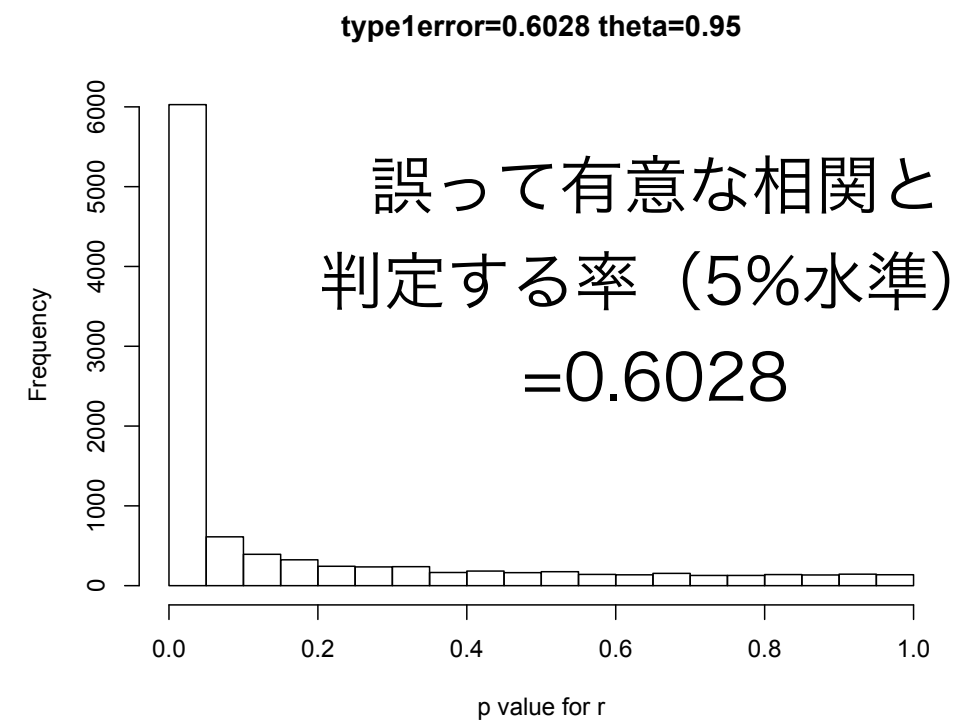
定常過程 θ_x の絶対値 < 1

$\theta_x = 1$ ランダムウォーク

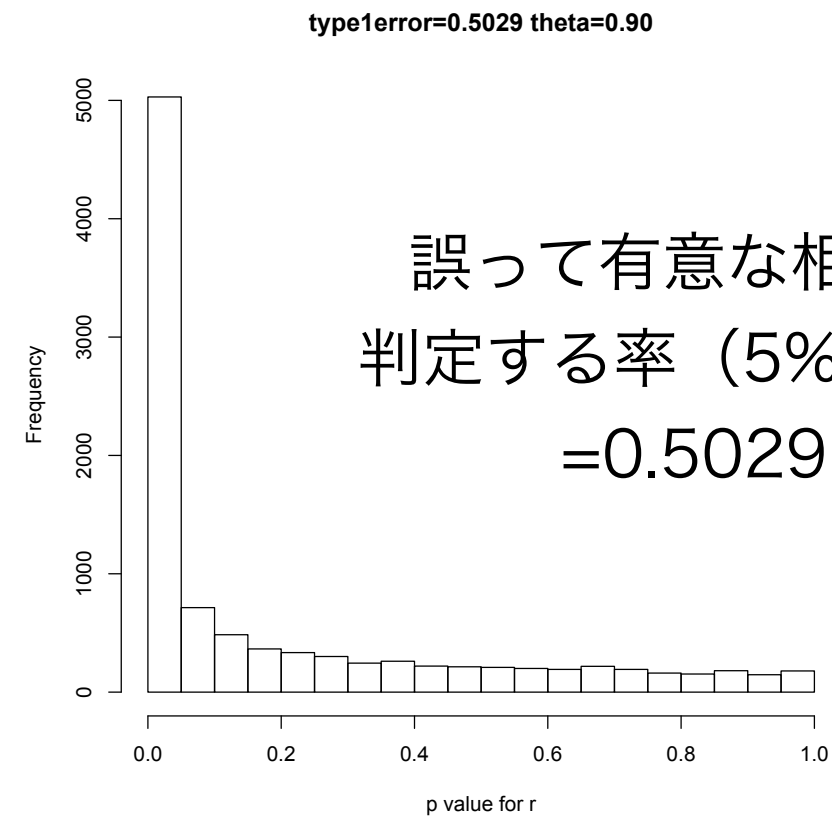
$$\theta x=0.98$$



$$\theta x=0.95$$



$$\theta x=0.90$$



回帰

回帰式 $y_t = \alpha + \beta x_t$

回帰モデルとしては

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

ε_t は平均0で分散一定の正規分布にしたがう

時系列データを回帰する（してしまう）とどうなる

回帰

回帰式 $y_t = \alpha + \beta x_t$

回帰モデルとしては

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

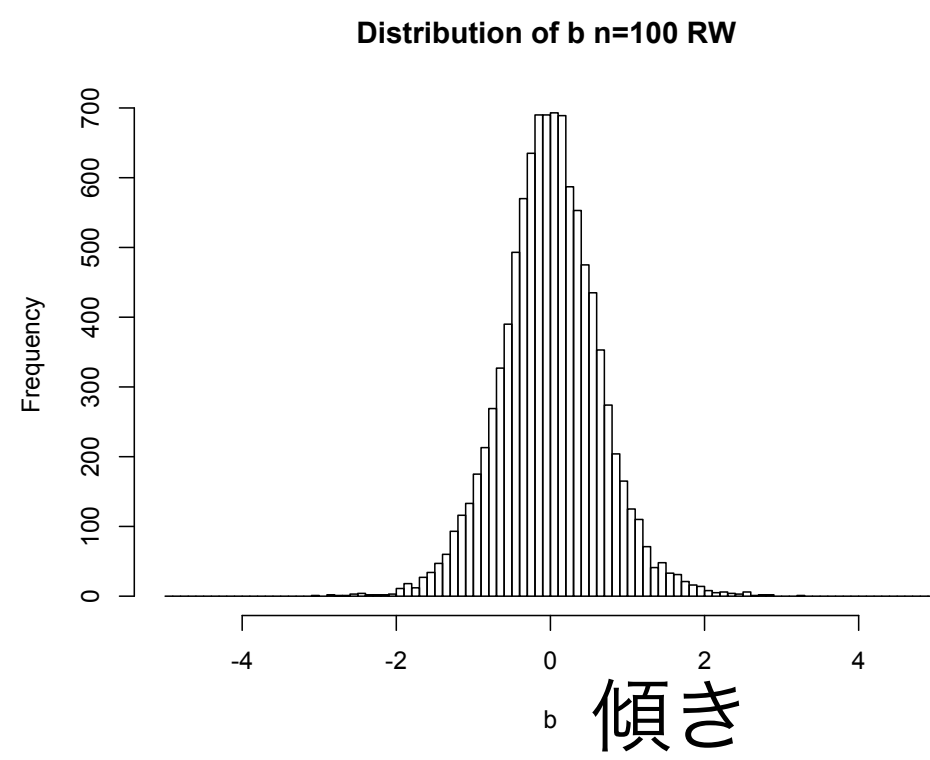
ε_t は平均0で分散一定の正規分布にしたがう

正規分布からの独立なデータ（回帰分析が仮定している条件）

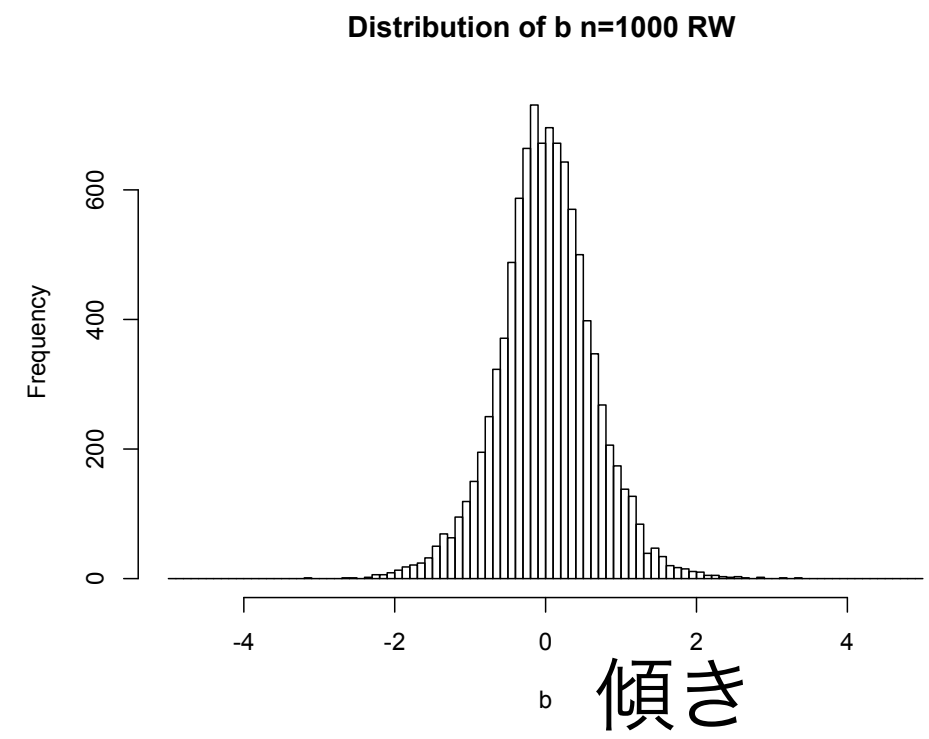
**パラメーター α 、 β のばらつきは、
サンプルサイズが大きいと、小さくなる**

傾きのヒストグラム

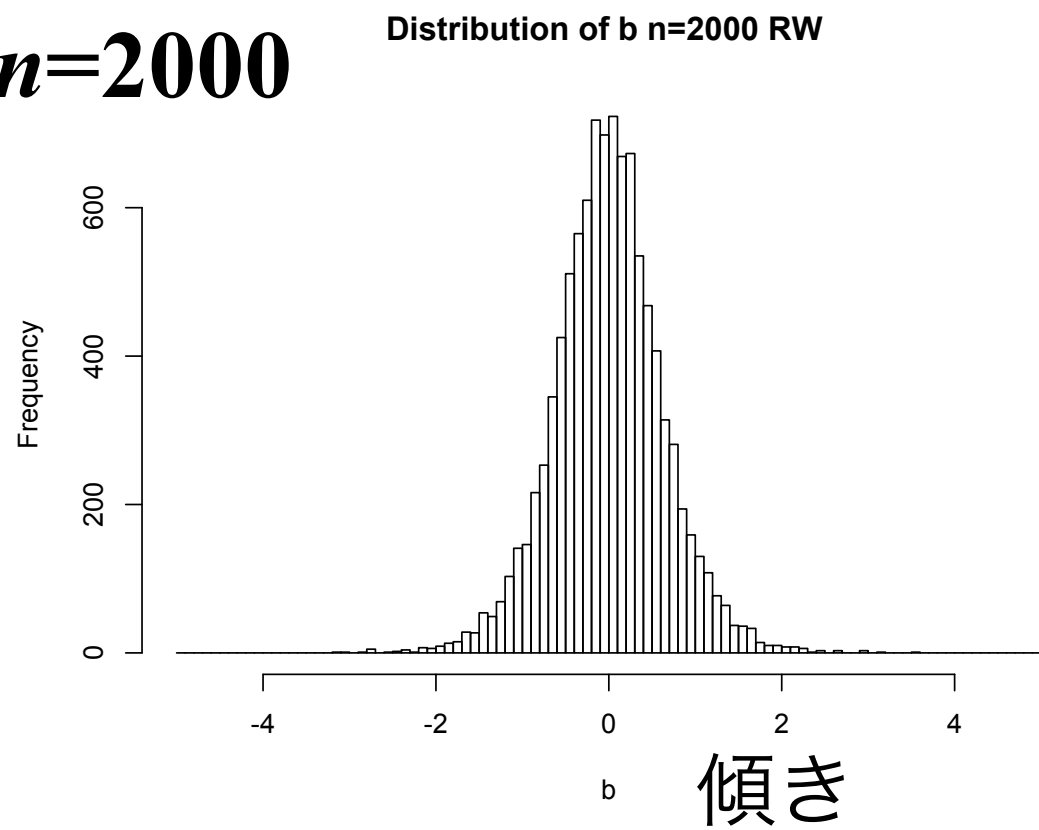
$n=100$



$n=1000$



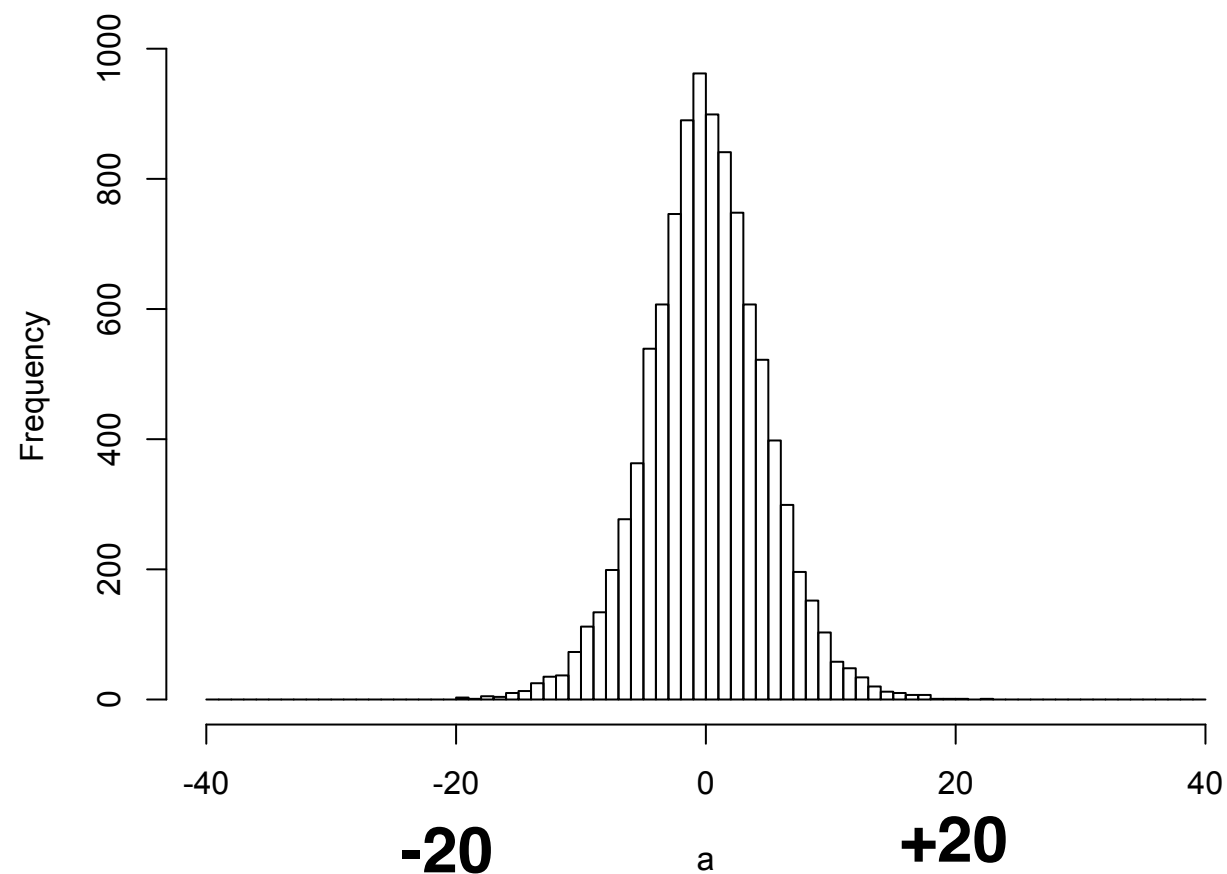
$n=2000$



切片

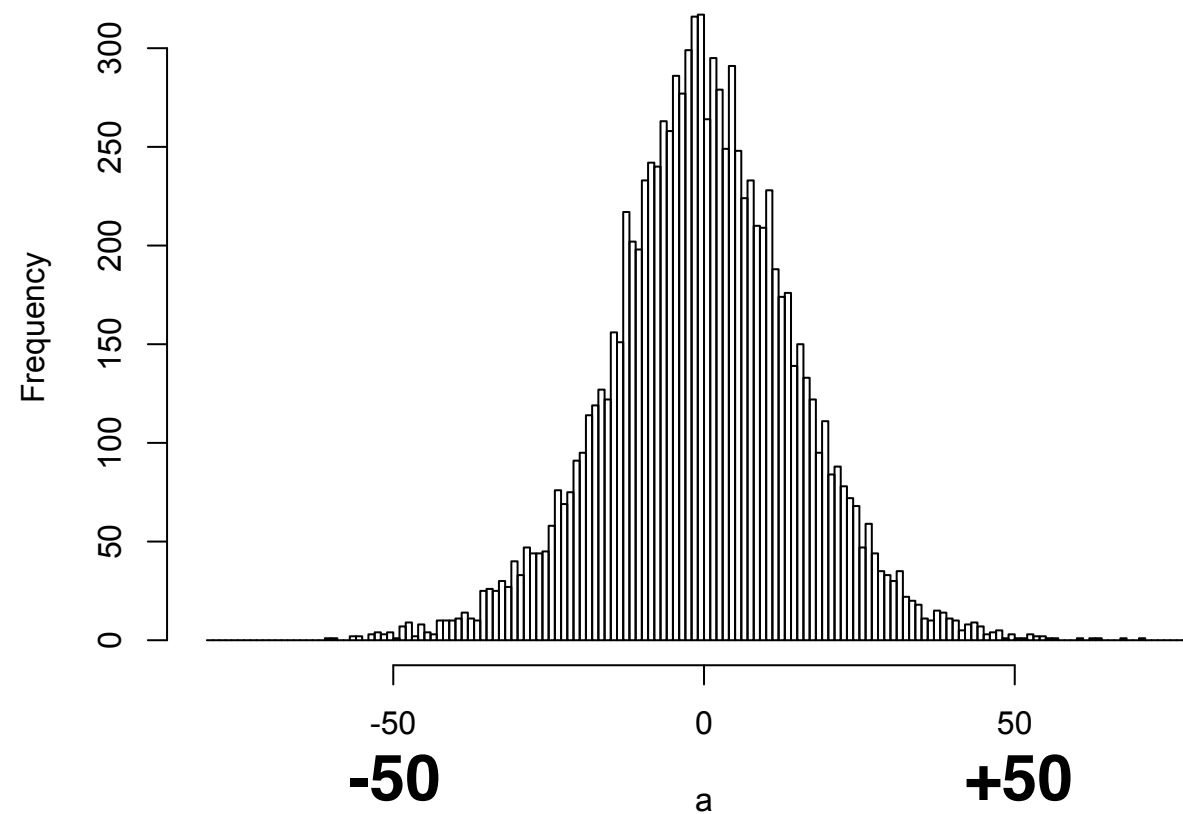
$n=100$

Distribution of a $n=100$ RW



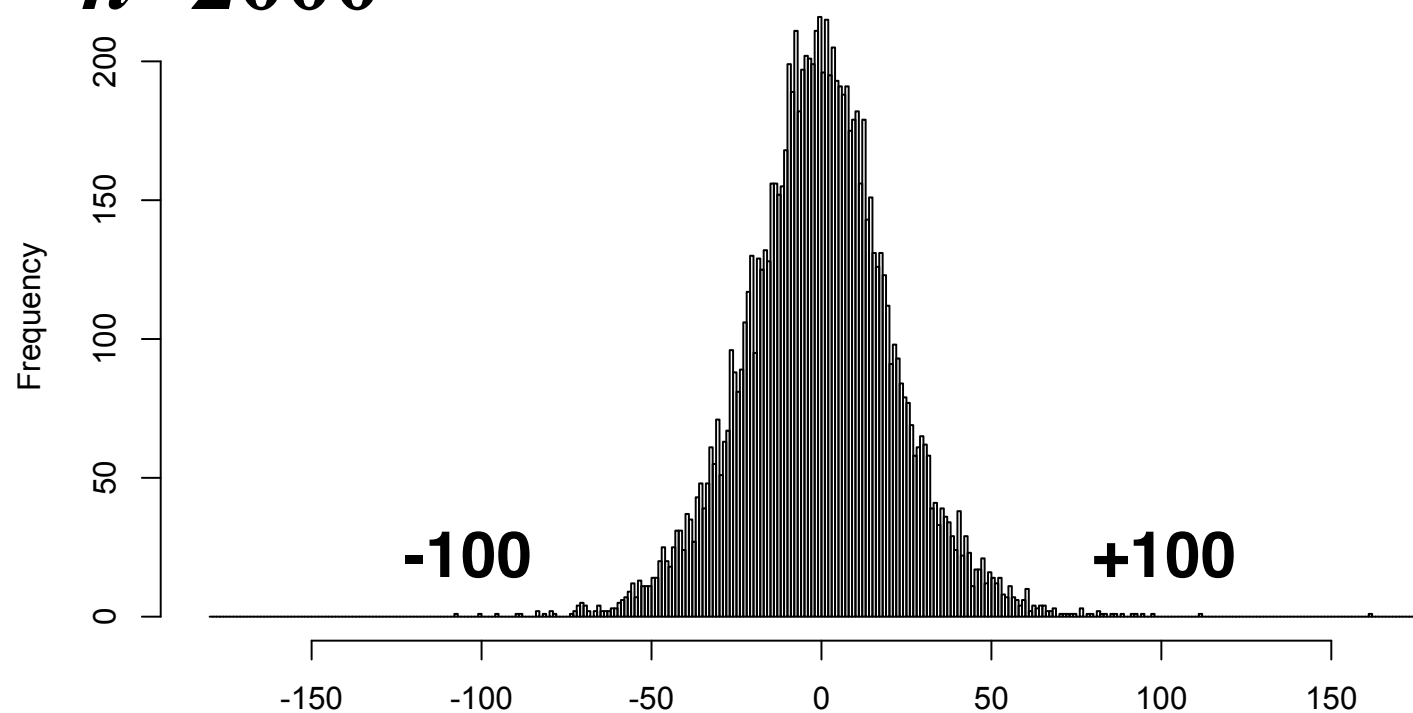
$n=1000$

Distribution of a $n=1000$ RW

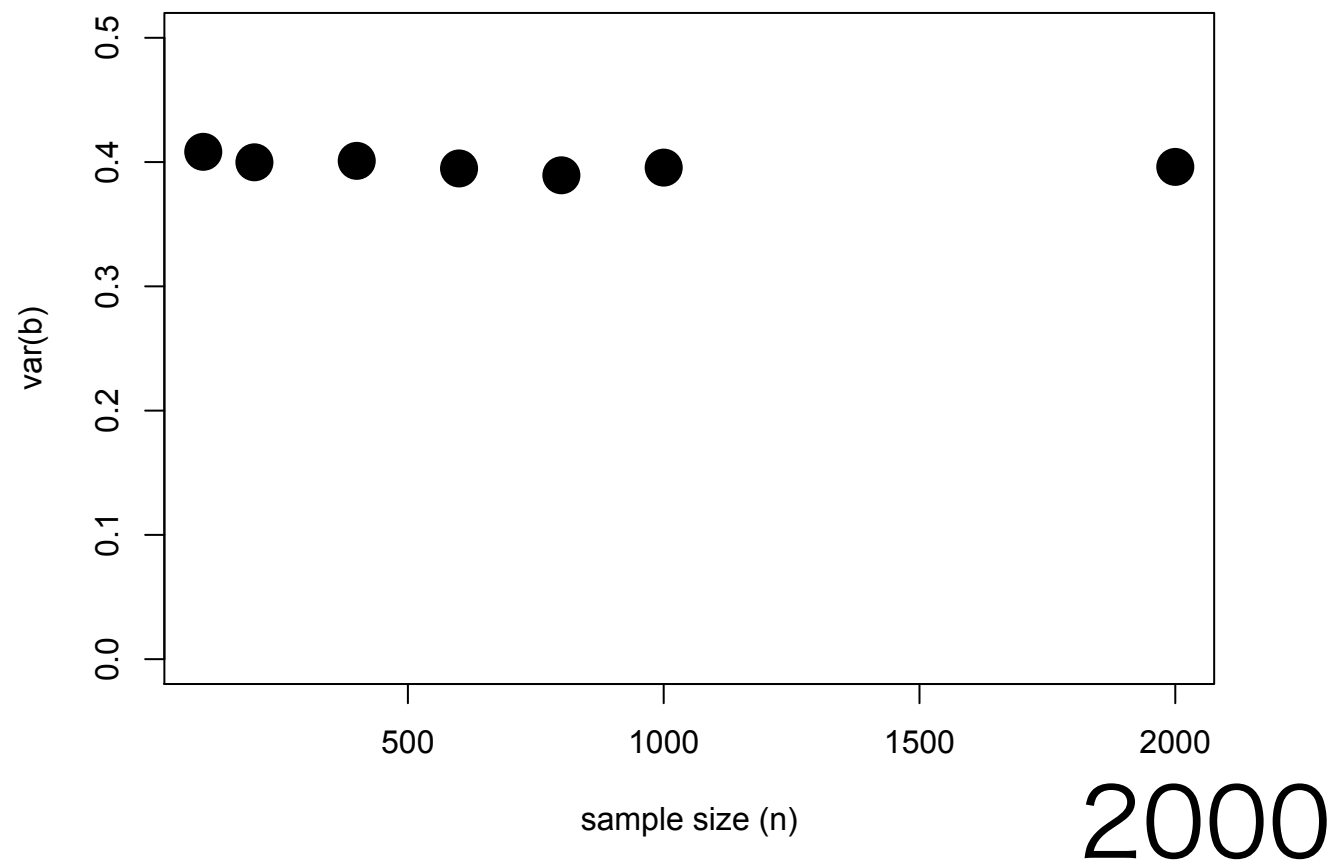


Distribution of a $n=2000$ RW

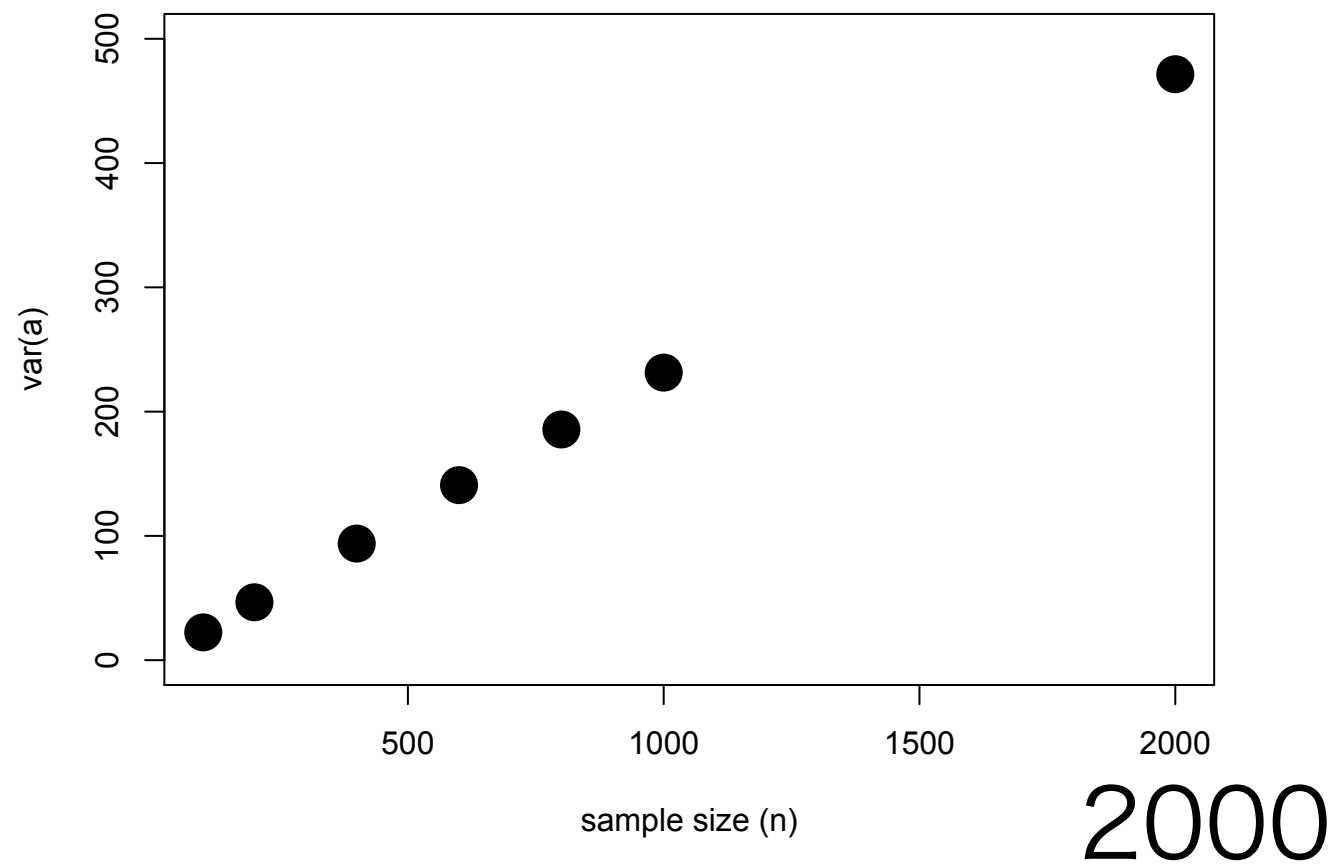
$n=2000$



傾きの分散



切片の分散



サンプル・サイズ

時系列の長さ

2つのランダムウォーク（互いに独立）

回帰すると、変な結果

見せかけの回帰 spurious regression

検定：第1種の誤りが過大

サンプル・サイズが大きくなるとさらに悪化

係数の推定値

傾き：サンプル・サイズが増えても、ばらつきが減らない

切片：サンプル・サイズが増えると、ばらつきが増える

まとめ

時系列

今回は”ランダムウォークごとき”

似たデータが続けて出やすい

時系列データをそのまま使うと、
そこは、相関も回帰も、常識的な働きをしない

魔界

付. こんな場合には

x ランダム・ウォーク

y

独立な正規乱数

各時点の資源量 + 新たに得た量

閾値を超えたら、放出して次時点の資源量はゼロになる

閾値以下なら、放出せず、

次時点の資源量は各時点の資源量 + 新たに得た量

y 放出された量

x ランダム・ウォーク

各時点で、新たに資源を受け取る——独立な正規乱数

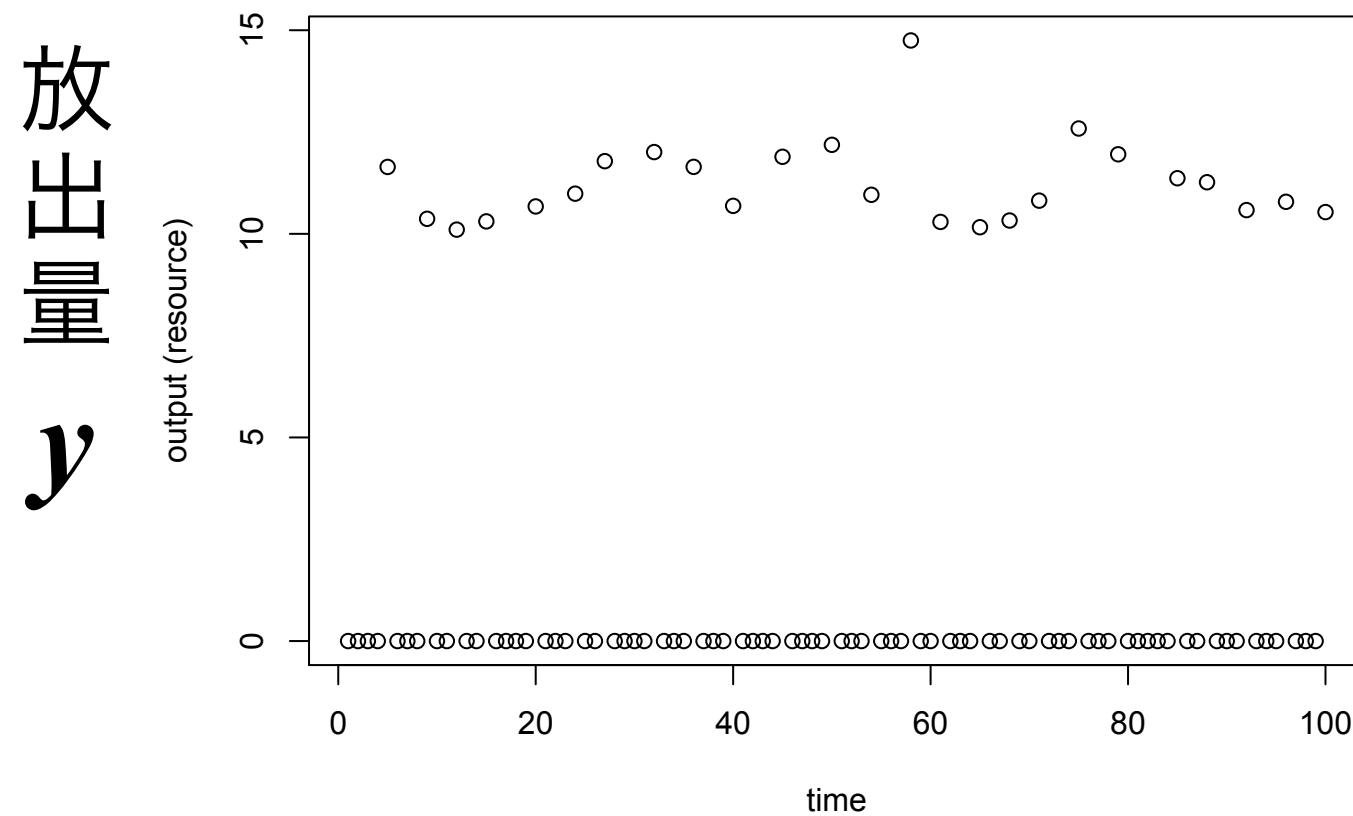
手持ちの資源量が閾値を超えたら、全部、放出する
放出すると次時点の資源量はゼロになる

手持ちの資源量が閾値以下なら、放出せず、
次時点の資源量は各時点の資源量＋新たに得た量

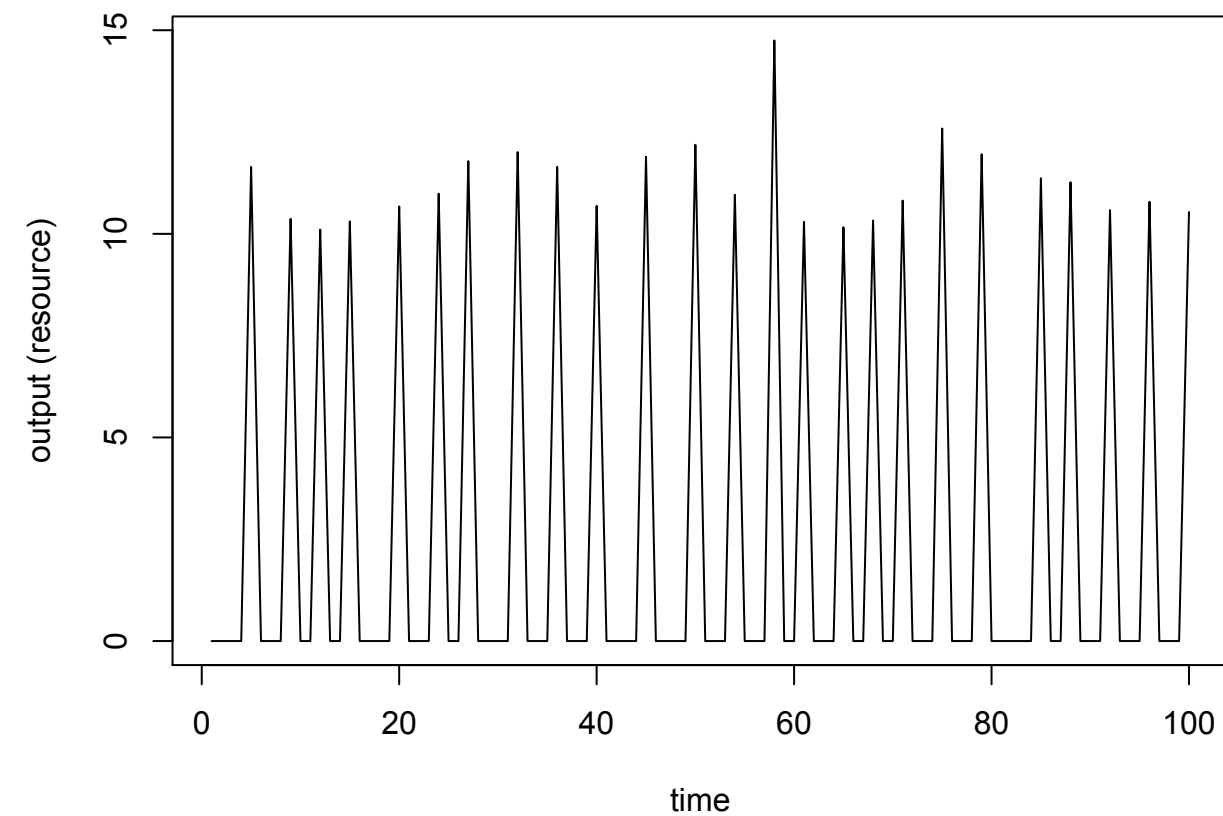
y 放出された量

例

resource output time series

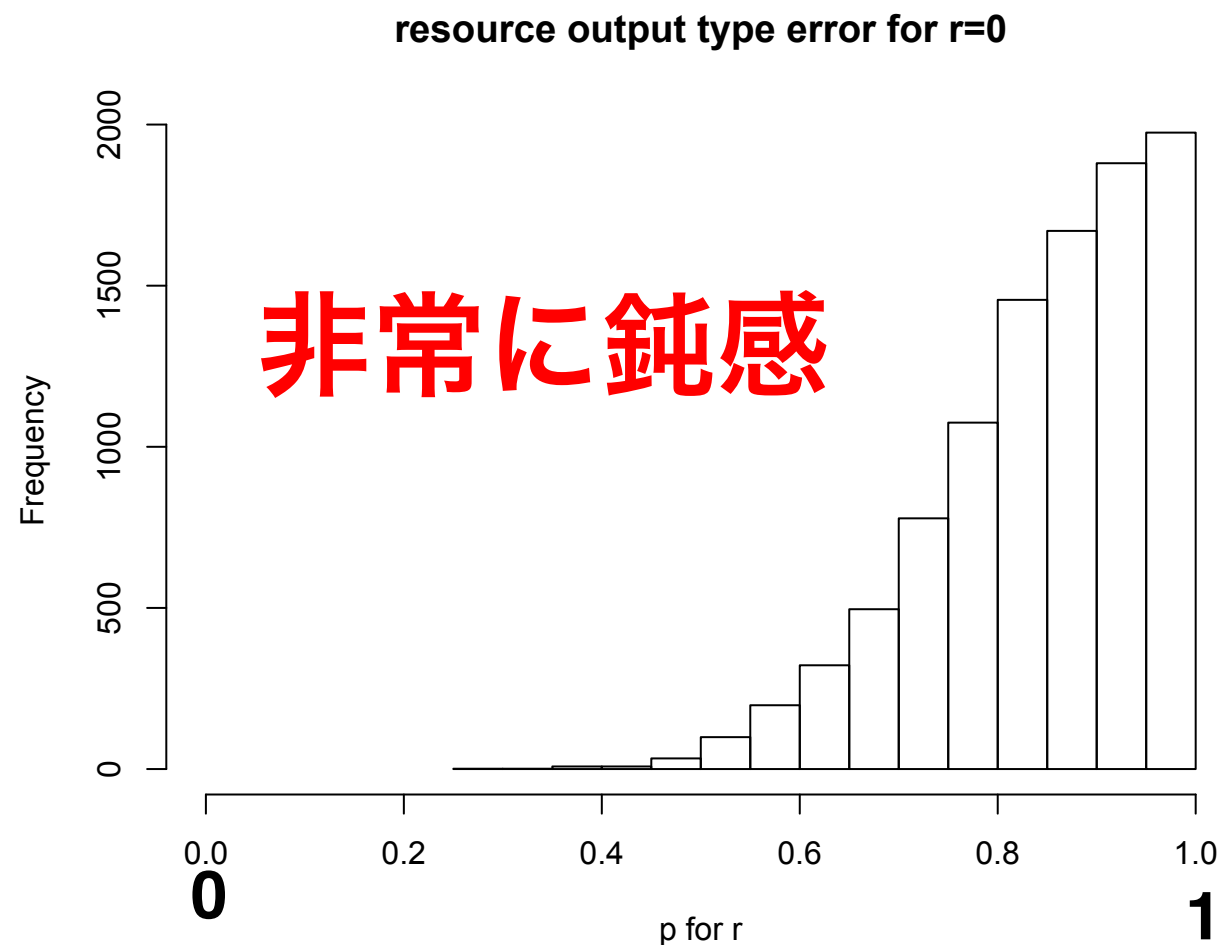


resource output time series



左右のグラフとも、同一の例（グラフの描き方がちがうだけ）

x と y の間の相関係数を
検定した、p の値の分布



p の値 (有意確率)

x と y の間の相関係数の分布

