シミュレーション工学

モンテカルロ法(2) 誤差解析と不偏推定量

慶應義塾大学大学院理工学研究科基礎理工学専攻物理情報専修

渡辺宙志

はじめに

測定と誤差

- 一般に測定値には実験誤差がある
- 数値計算においても、測定結果は誤差を伴う
- 「誤差」の理解は難しい

本講義の目的

- エラーバーとは何か、どのような性質を持つかを理解する
- 統計誤差と系統誤差について理解する
- 系統誤差を除去する(Jackknife法)

エラーバーとは

ある回路の電流を6回測定したら、以下のデータ[mA]を得た

1.16, 1.13, 1.12, 1.12, 1.11, 1.08

この回路の電流の観測値は?

観測値　　エラーバー

$$1.12 \pm \boxed{0.01}\,\mathrm{mA}$$

大雑把な意味：観測値の1.1までは自信があるが、小数点第二位は自信がなく、1.11かもしれないし1.13かもしれない

正確な定義は？

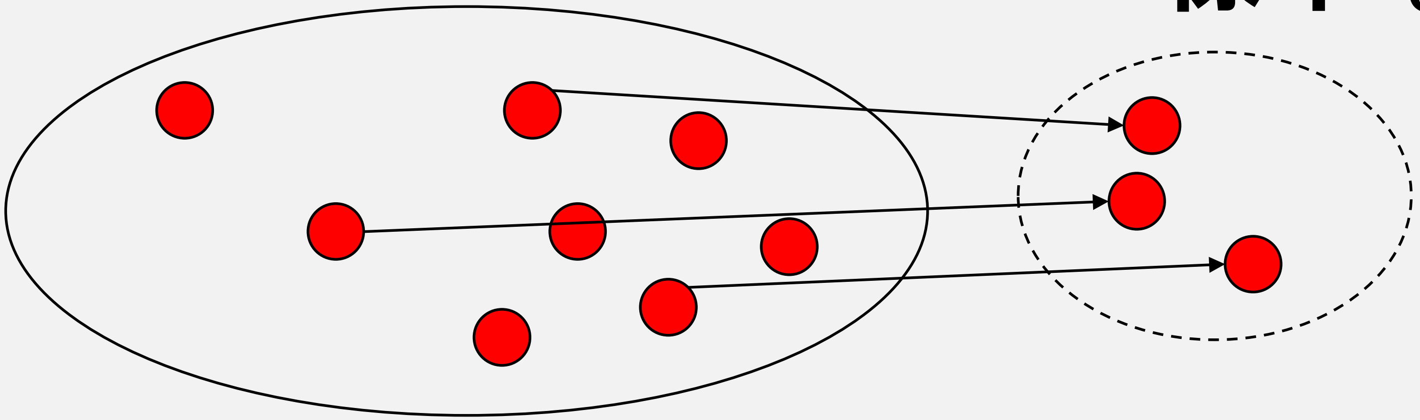
エラーバーとは

観測するたびに値が変化する量を確率変数$\hat{X}$とみなす
この変数の「全ての可能性の集合」を**母集団**と呼ぶ

観測により母集団から**標本**を取り出す

母集団 $\{\hat{X}\}$

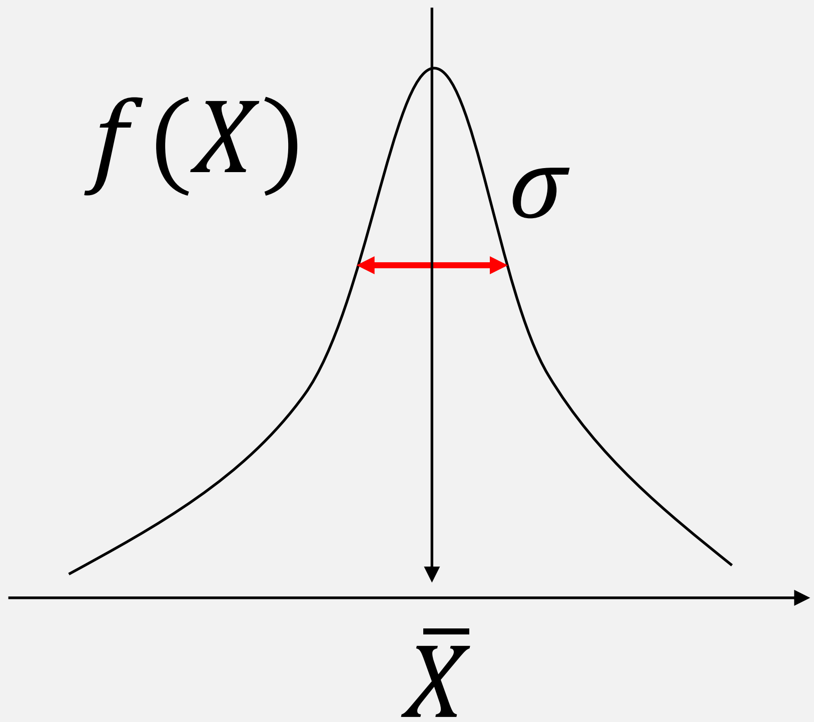
標本 $\{X_i\}$



標本の集合から**母集団の性質**を**推定**するのが目的

エラーバーとは

母集団の特徴量(平均や分散)を知りたい



分布の1次のモーメント

$$\bar{X} = \int X f(X) dX$$

分布の2次のモーメント

$$\sigma^2 = \int (X - \bar{X})^2 f(X) dX$$

手元にあるのはN個の標本$\{X_i\}$
標本から特徴量を得る関数を推定量(estimator)と呼ぶ

エラーバーとは

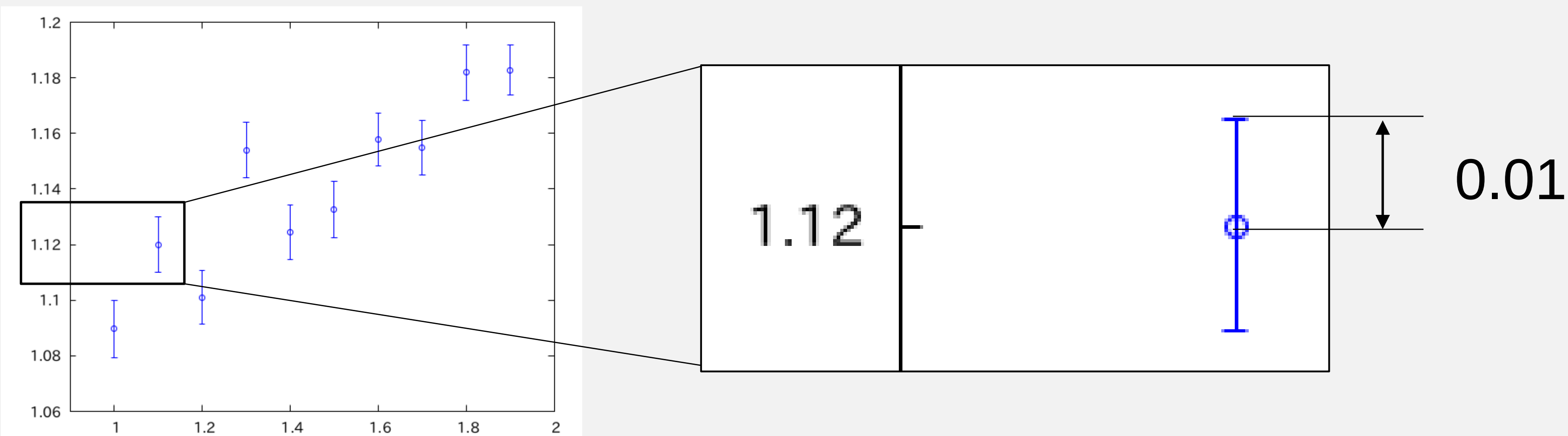
推定したい量	そのestimator
平均値	$\bar{X} = \frac{1}{N}\sum_{i}^{N} X_i$
母分散	$\sigma^2 = \frac{1}{N-1}\sum_{i}^{N}(X_i - \bar{X})^2$
平均値の推定値の分散	$\sigma_{\bar{X}}^2 = \frac{1}{N(N-1)}\sum_{i}^{N}(X_i - \bar{X})^2$

※ NではなくN-1で割るのは不偏分散を求めるため

エラーバーとは

平均値の推定値の標準偏差を誤差とみなす

$$\bar{X} \pm \sqrt{\sigma_{\bar{X}}^2} \qquad 1.12 \pm 0.01\ \mathrm{mA}$$

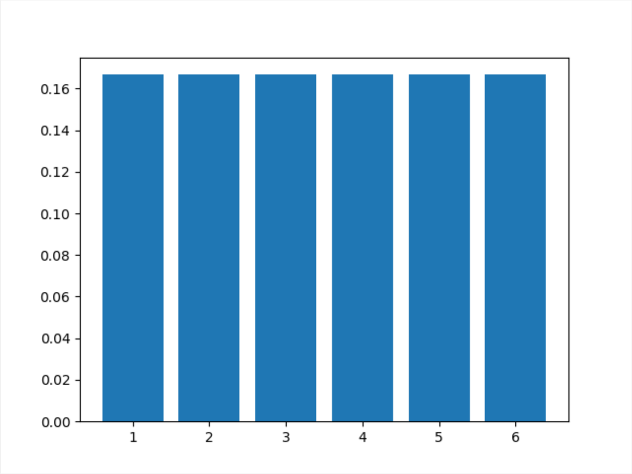


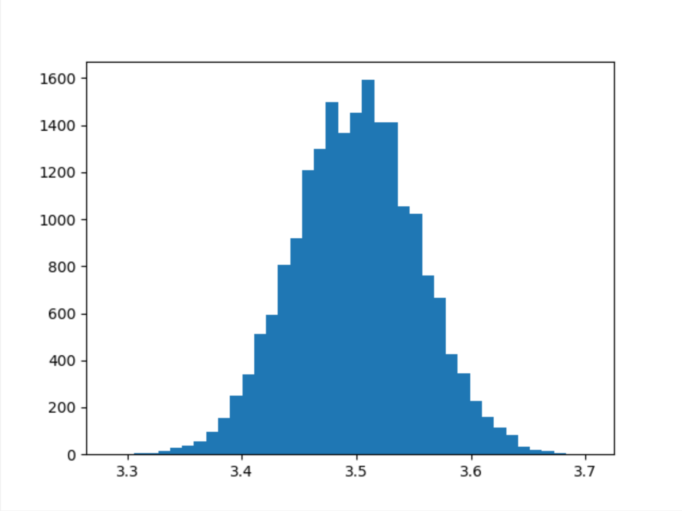
平均値の推定値の分散の平方根をエラーバーとする

エラーバーの意味は？

中心極限定理

- 一般に、観測値の分布はガウス分布ではない
- しかし、観測値が**独立同分布**に従う確率変数とみなせる時、その期待値は**ガウス分布**に近づく



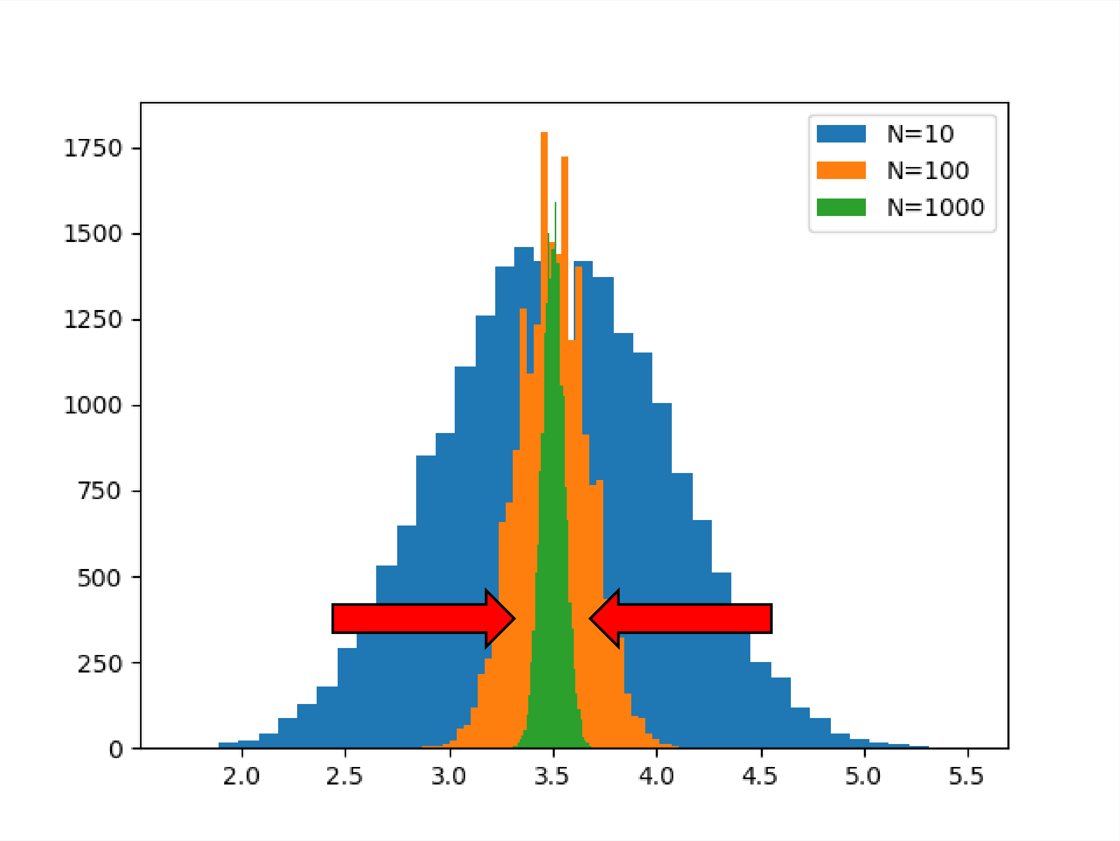


サイコロの目の**母集団の分布**は一様分布だが、

千回振った目の**平均値の分布**はガウス分布に近づく

中心極限定理

標本が多くなるほど平均値の分布の分散は小さくなる



N=10, 100, 1000回サイコロを振った時、出た目の平均の頻度分布

中心極限定理

サンプル数Nを増やした時

母分散のestimatorは一定値に収束する

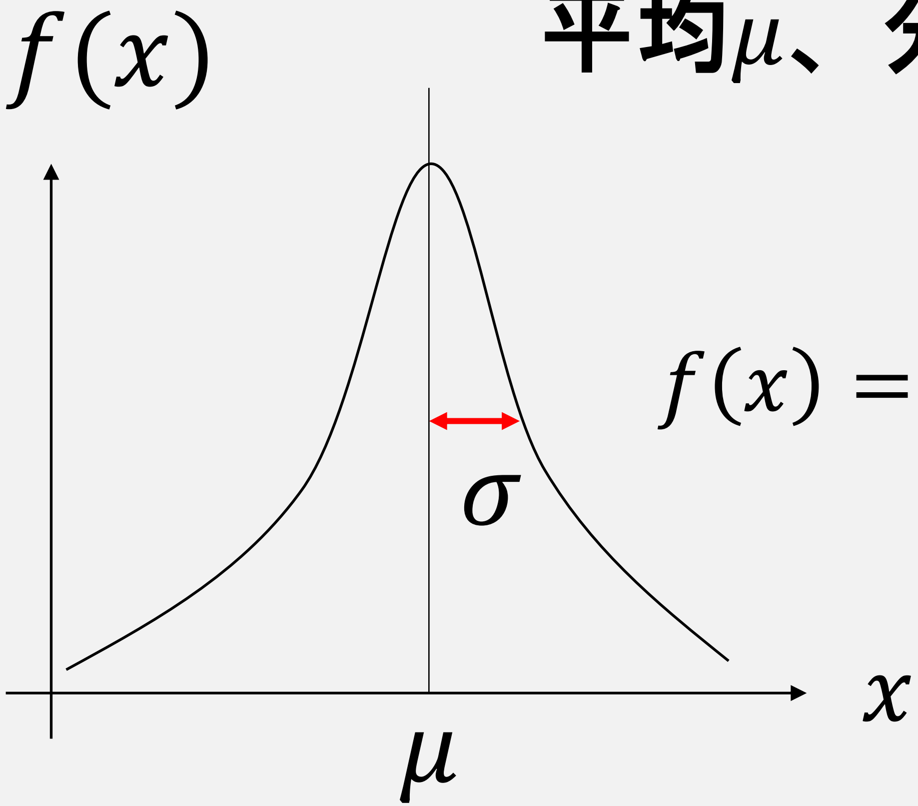
$$\sigma^2 = \frac{1}{N-1}\sum_i^N (X_i - \bar{X})^2 \qquad \lim_{N\to\infty} \sigma^2 = \text{const.}$$

平均値の推定値の分散のestimatorは$1/N$の早さでゼロに収束する

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N} \qquad \lim_{N\to\infty} \sigma_{\bar{X}}^2 = 0$$

ガウス分布

平均μ、分散σ^2のガウス分布



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

平均μ：分布の中心の位置
標準偏差σ：分布の幅

ガウス分布

確率変数$\hat{X}$の値がaとbの間にある確率が

$$P\left(a < \hat{X} < b\right) = \int_a^b f(x)dx$$

で与えられる時、$f(x)$を$\hat{X}$の**確率密度関数**と呼ぶ

確率密度関数が平均μ、分散σ^2のガウス分布である時

$$\mu - \sigma < x < \mu + \sigma$$

の範囲を**1シグマの範囲**と呼ぶ

ガウス分布

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$ であるとき、

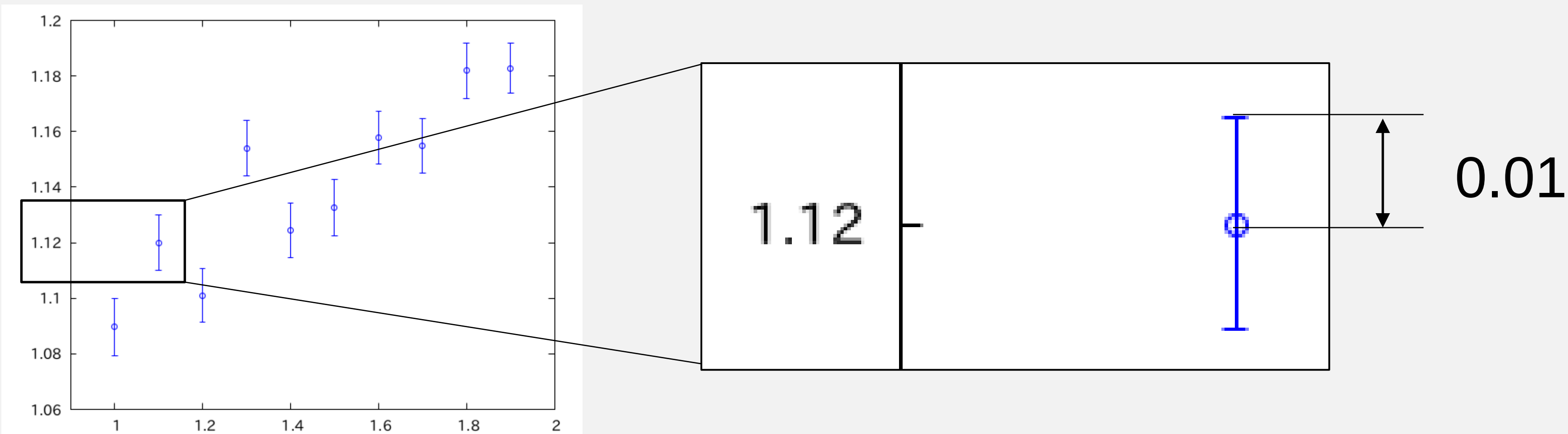
$$P\left(\mu - \sigma < \hat{X} < \mu + \sigma\right) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)dx \sim 0.6827$$

平均μ、分散σ^2のガウス分布に従う確率変数が、
平均の周りにσの間で揺らぐ確率が68.27%

ex) **テストで偏差値**40〜60**までの間の人が**68.27%

ガウス分布

エラーバーを平均値の推定値の標準偏差とする(1シグマの範囲)



同様な実験を繰り返した場合、観測値がエラーバーの間に入る確率が68.27%

ガウス分布

同様に「nシグマの範囲」が定義できる

$$\mu - n\sigma < x < \mu + n\sigma$$

1シグマ：入る確率 68.27%　外れる確率 31.73%
2シグマ：入る確率 95.45%　外れる確率 4.55%
3シグマ：入る確率 99.73%　外れる確率 0.27%
5シグマ：入る確率 99.9994%　外れる確率 0.0005%

エラーバーのまとめ

- エラーバーとは観測値を確率変数とみなした時に、その平均値の分布の推定標準偏差のこと
- サンプル数を増やせば増やすほど、エラーバーは小さくなる
- 観測値が独立同分布なら、サンプル数を増やしていくと平均値の分布はガウス分布に漸近する
- 平均μ、分散σ^2のガウス分布について、以下を「nシグマの範囲」と呼ぶ

$$\mu - n\sigma < x < \mu + n\sigma$$

- ガウス分布に従う確率変数が独立であるなら
  - 「1シグマの範囲」からは3つに1つは外れる
  - 「5シグマの範囲」から外れる確率はほぼゼロ

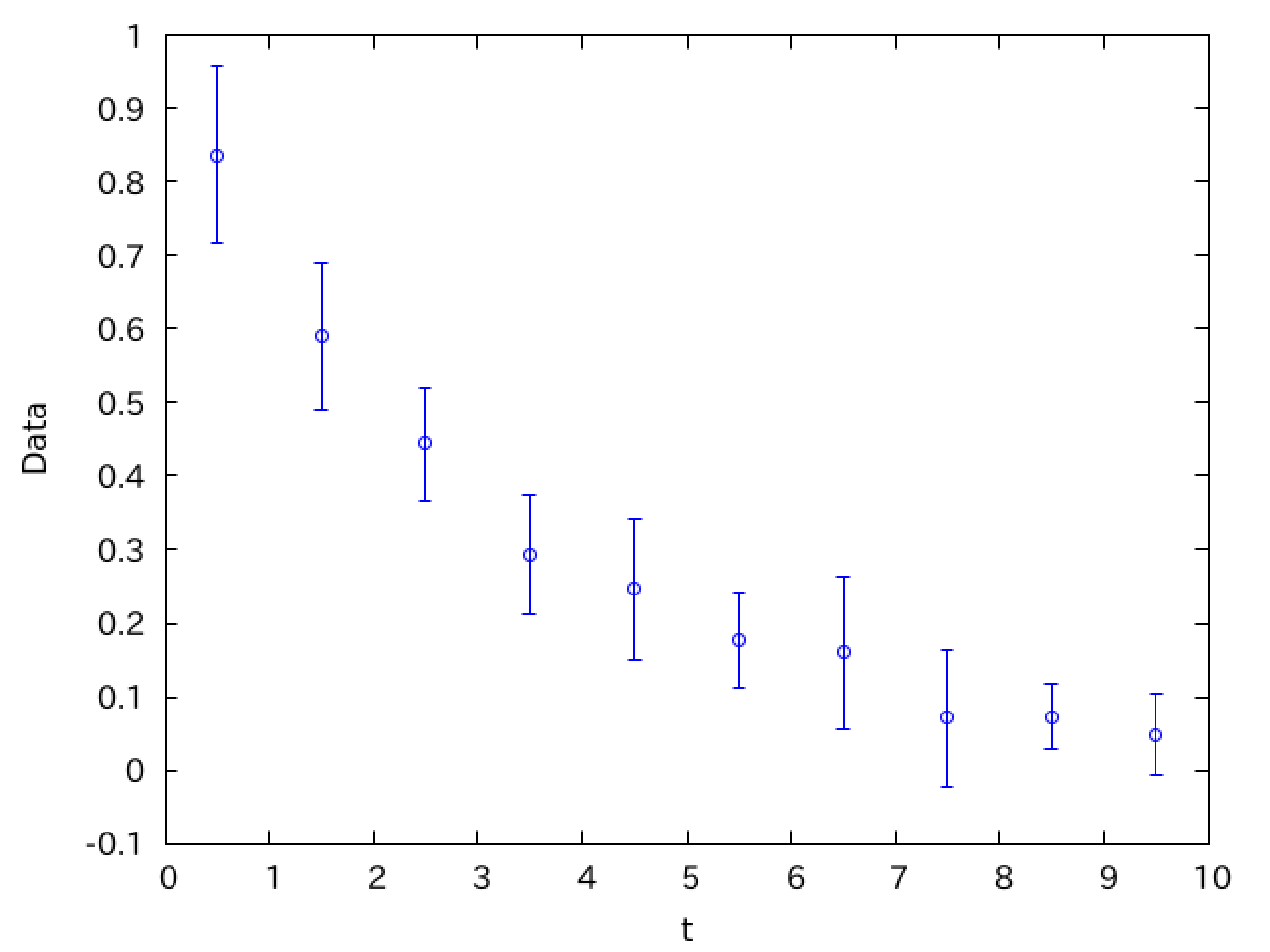
エラーバーの活用

データがガウス分布に従い、かつ独立であるとする観測量の母集団の分布の平均を「真の値」と呼ぶと

- 観測値は「真の値」の上下に均等にばらつく
- 観測値の3つに1つが「真の値」の1シグマの範囲に入らない
- 観測値と「真の値」がエラーバーの2倍離れることは稀、5倍離れることはまずない

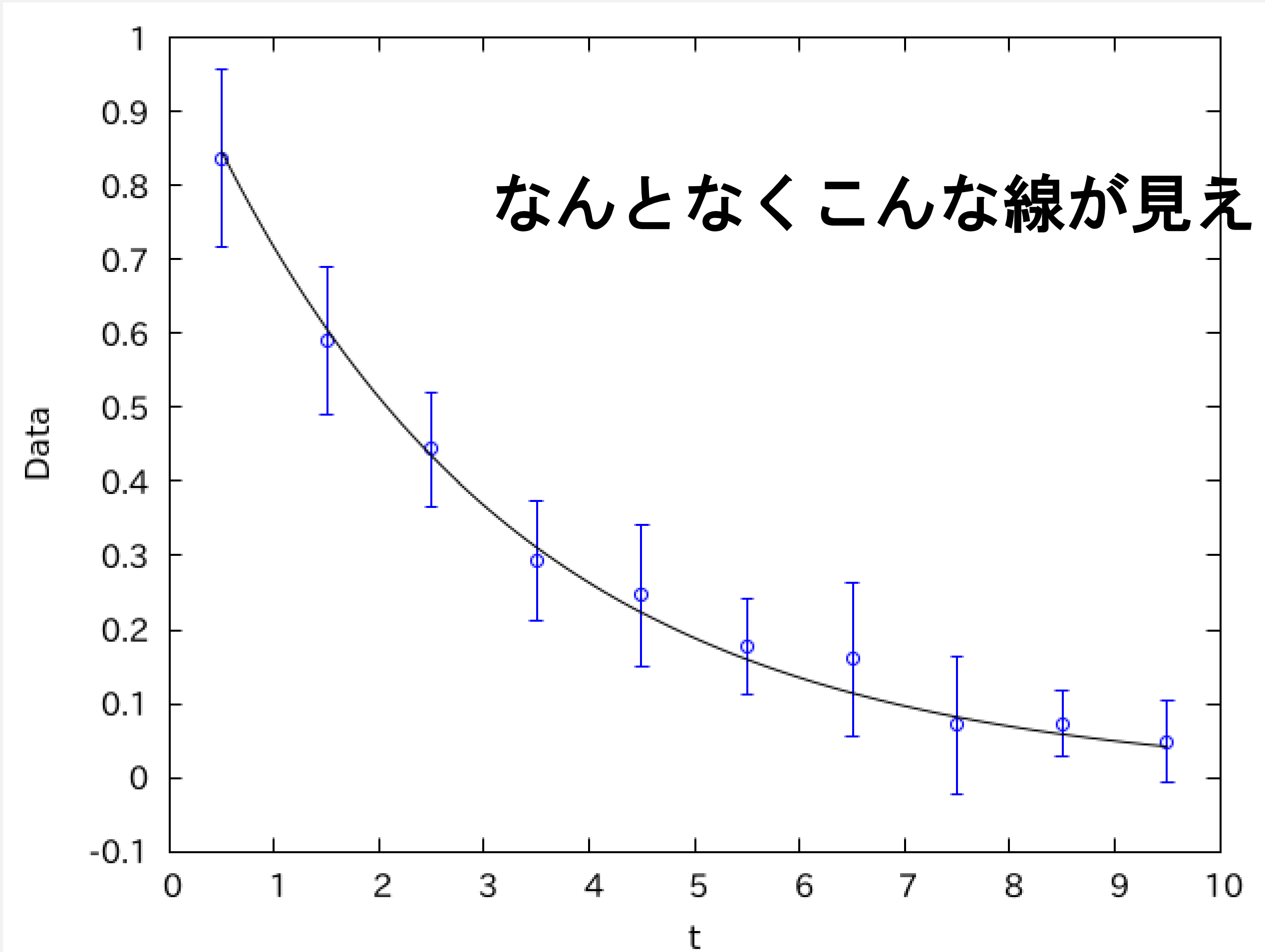
この知識を活用して「おかしなグラフ」に気づくことができる

おかしいグラフ1



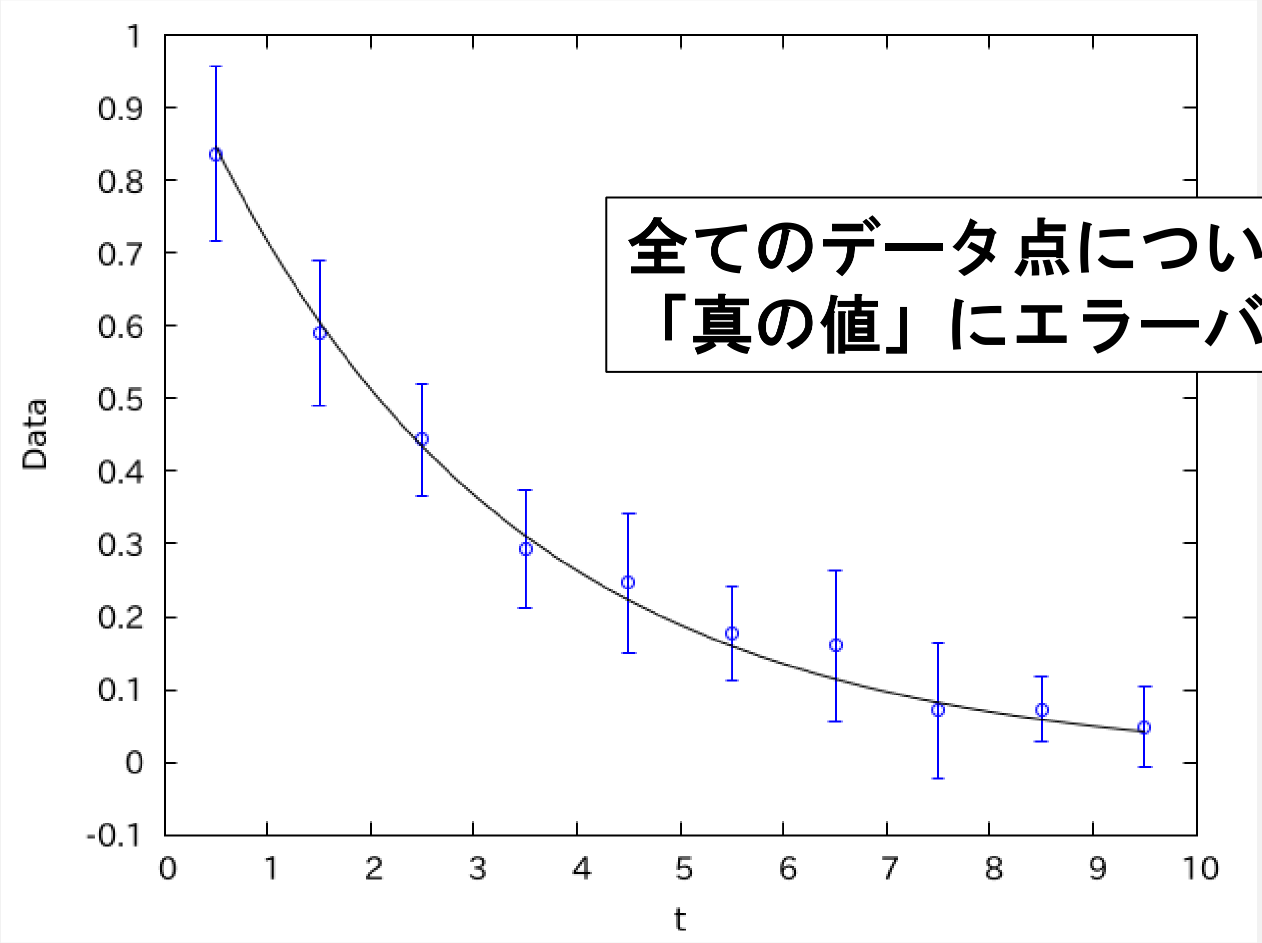
何かが指数関数的に減衰しているようだが・・・？

おかしいグラフ1



計算精度を高くしていったら、データはこの線に収束するであろうと期待される線→「真の値」

エラーバーがおかしいグラフ1



全てのデータ点について
「真の値」にエラーバーがかかっている

もしエラーバーが1シグマの範囲で取られていたら
3つに1つは「真の値」から外れないとおかしい

おかしいグラフ1

先ほどのデータを生成したコード

```
import numpy as np

N = 10
np.random.seed(1)
for i in range(10):
    x = i + 0.5
    d = np.zeros(N)
    d += np.exp(-x/3)
    d += np.random.randn(N)*0.1
    y = np.average(d)
    e = np.std(d)
    print(f"{x} {y} {e}")
```

エラーバーとしてnumpy.stdをそのまま使っている

平均値の推定誤差ではなく、母集団の標準偏差を求めてしまっている

おかしいグラフ1

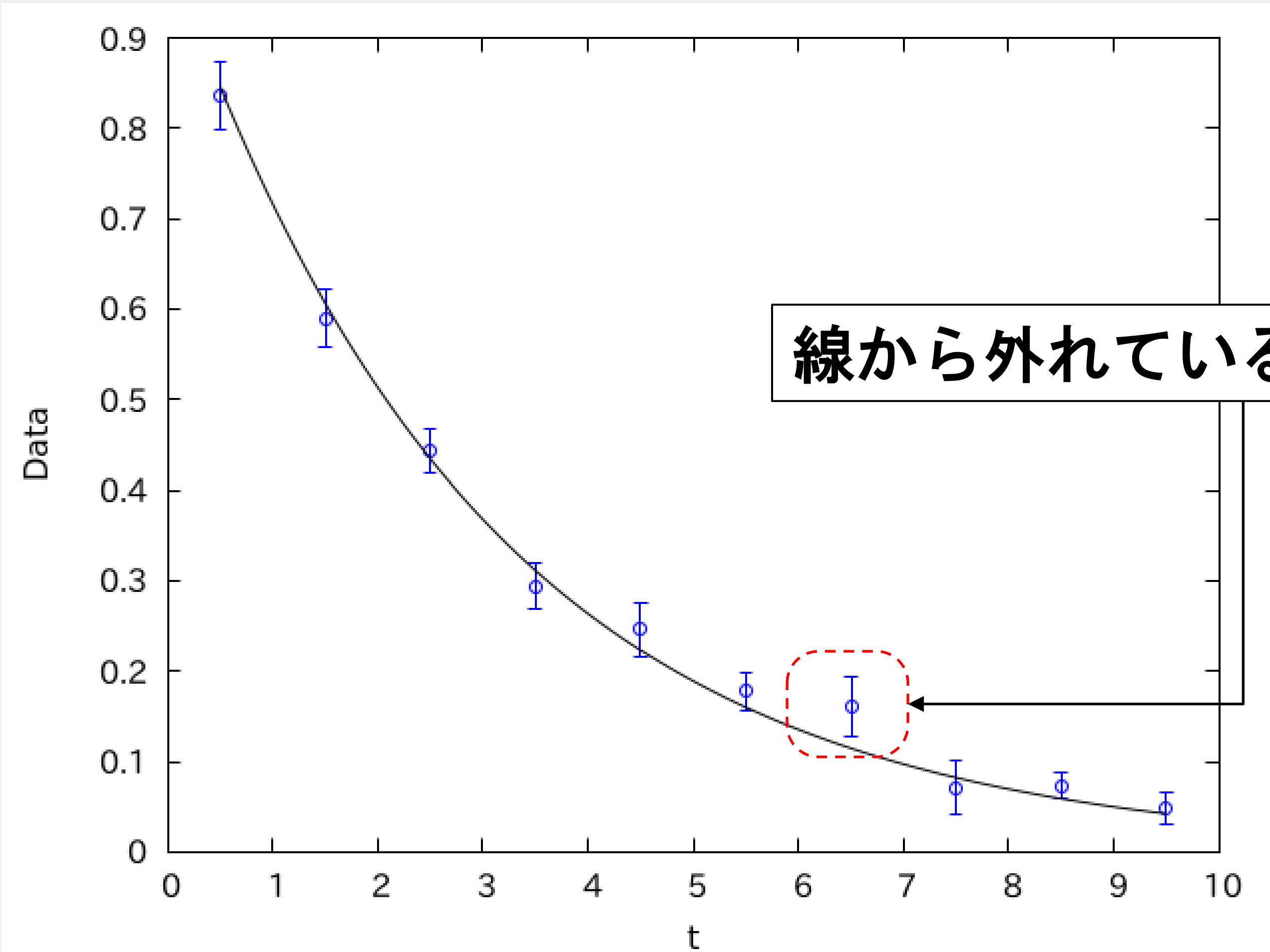
```python
import numpy as np

N = 10
np.random.seed(1)
for i in range(10):
    x = i + 0.5
    d = np.zeros(N)
    d += np.exp(-x/3)
    d += np.random.randn(N)*0.1
    y = np.average(d)
    e = np.std(d)/np.sqrt(N)
    print(f"{x} {y} {e}")
```

これが正しいコード

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

適切なグラフ

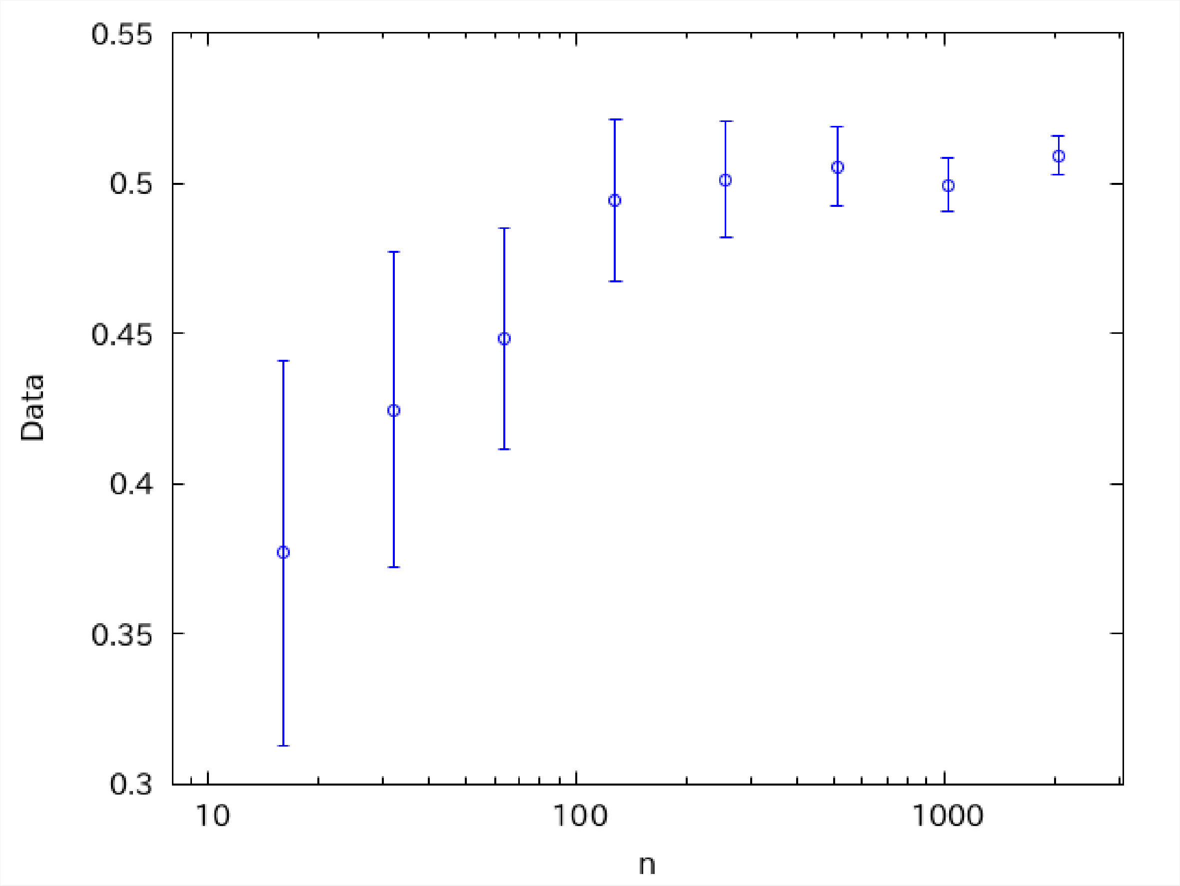


線から外れているデータがある

1シグマの範囲なら「外れているデータ」がないと不自然

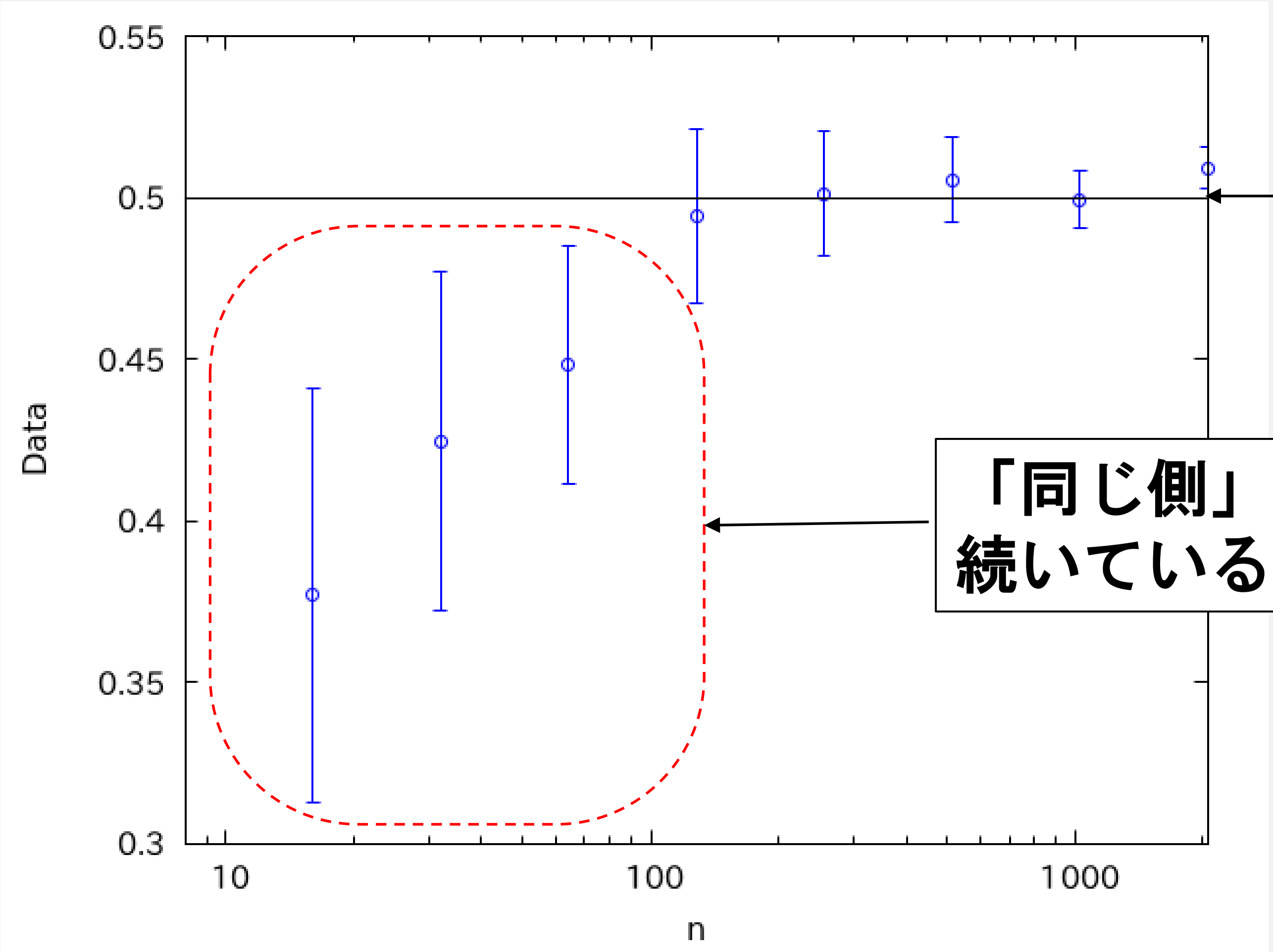
おかしいグラフ2

ある観測値のサンプル数n依存性のグラフ



サンプル数が増えると収束し、かつエラーバーが小さくなるのは
もっともらしいが・・・？

おかしいグラフ2



「真の値」はこのあたりにありそう

「同じ側」に外れてることが続いている

各データ点が独立なら、「真の値」の両側にばらつくはず

おかしいグラフ2

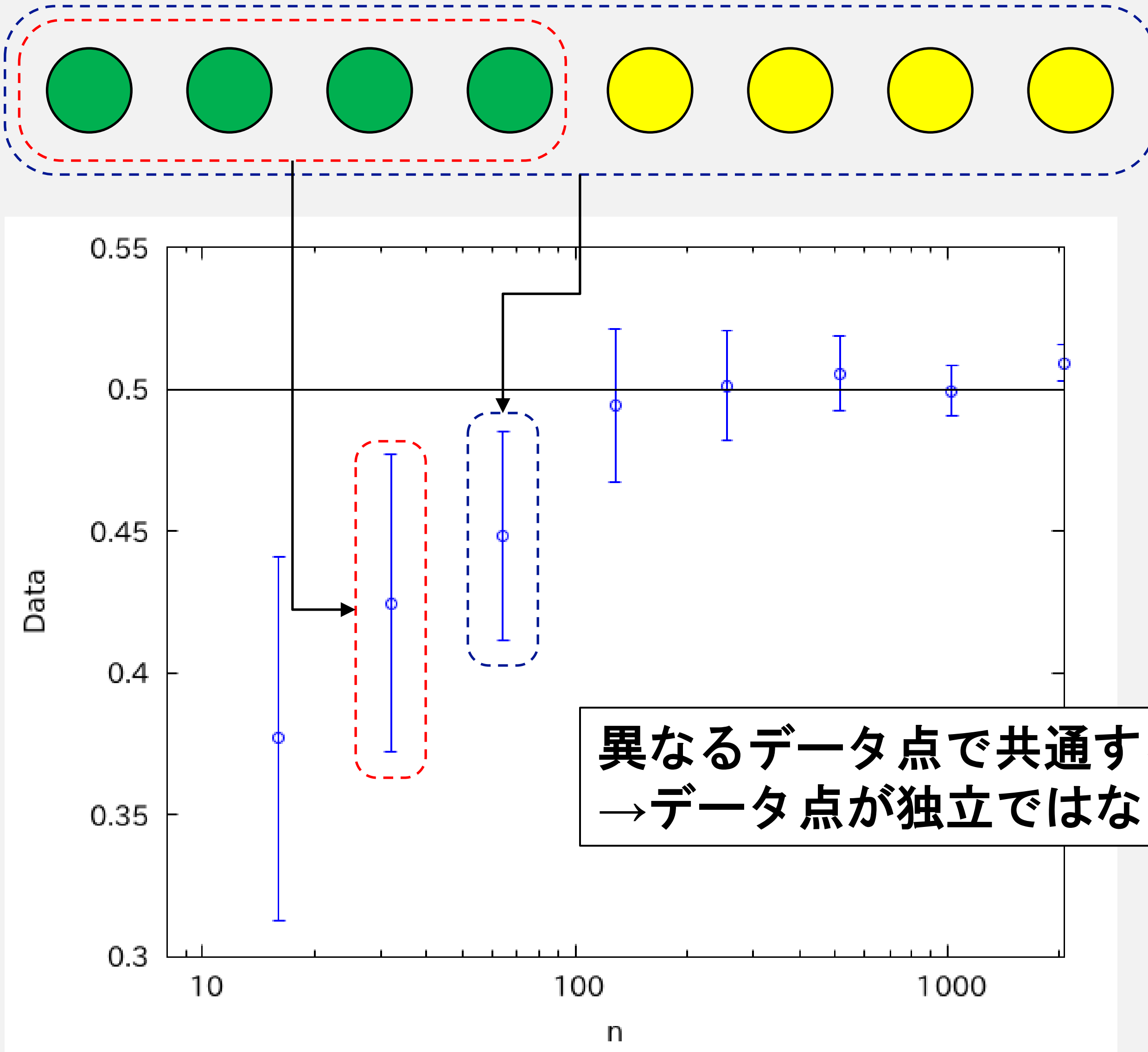
先ほどのデータを生成したコード

```
import numpy as np

np.random.seed(1)
N = 2048
d = np.random.random(N)
for i in range(4, 12):
    n = 2**i
    dd = d[:n]
    ave = np.average(dd)
    err = np.std(dd)/np.sqrt(n)
    print(f"{n} {ave} {err}")
```

先に全データを作成し、部分配列について誤差を計算している

おかしいグラフ2



異なるデータ点で共通するデータを使っている
→データ点が独立ではない

適切なグラフ

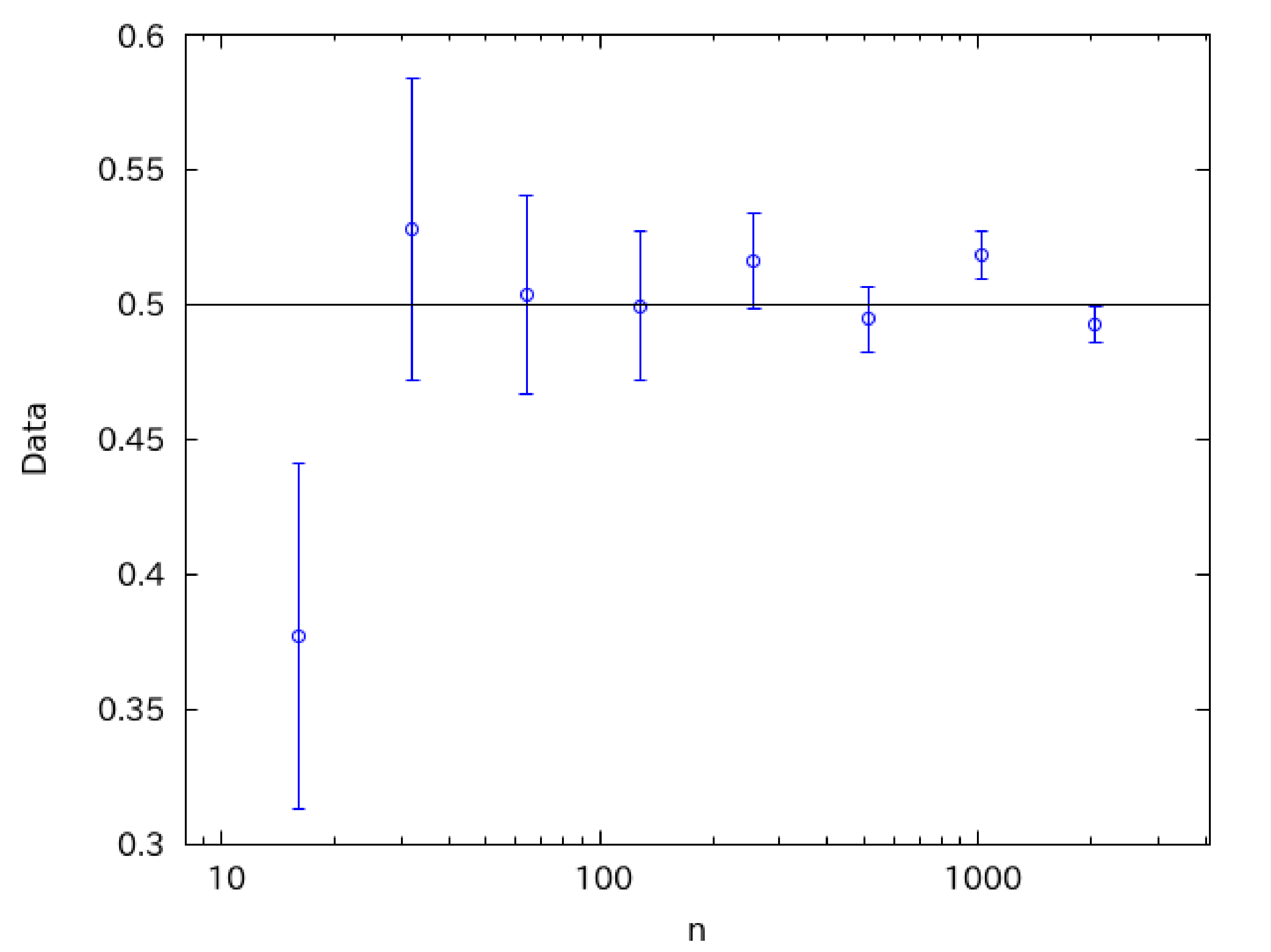
データを適切に生成するコード

```
import numpy as np

np.random.seed(1)
N = 2048
for i in range(4, 12):
    n = 2**i
    dd = np.random.random(n)
    ave = np.average(dd)
    err = np.std(dd)/np.sqrt(n)
    print(f"{n} {ave} {err}")
```

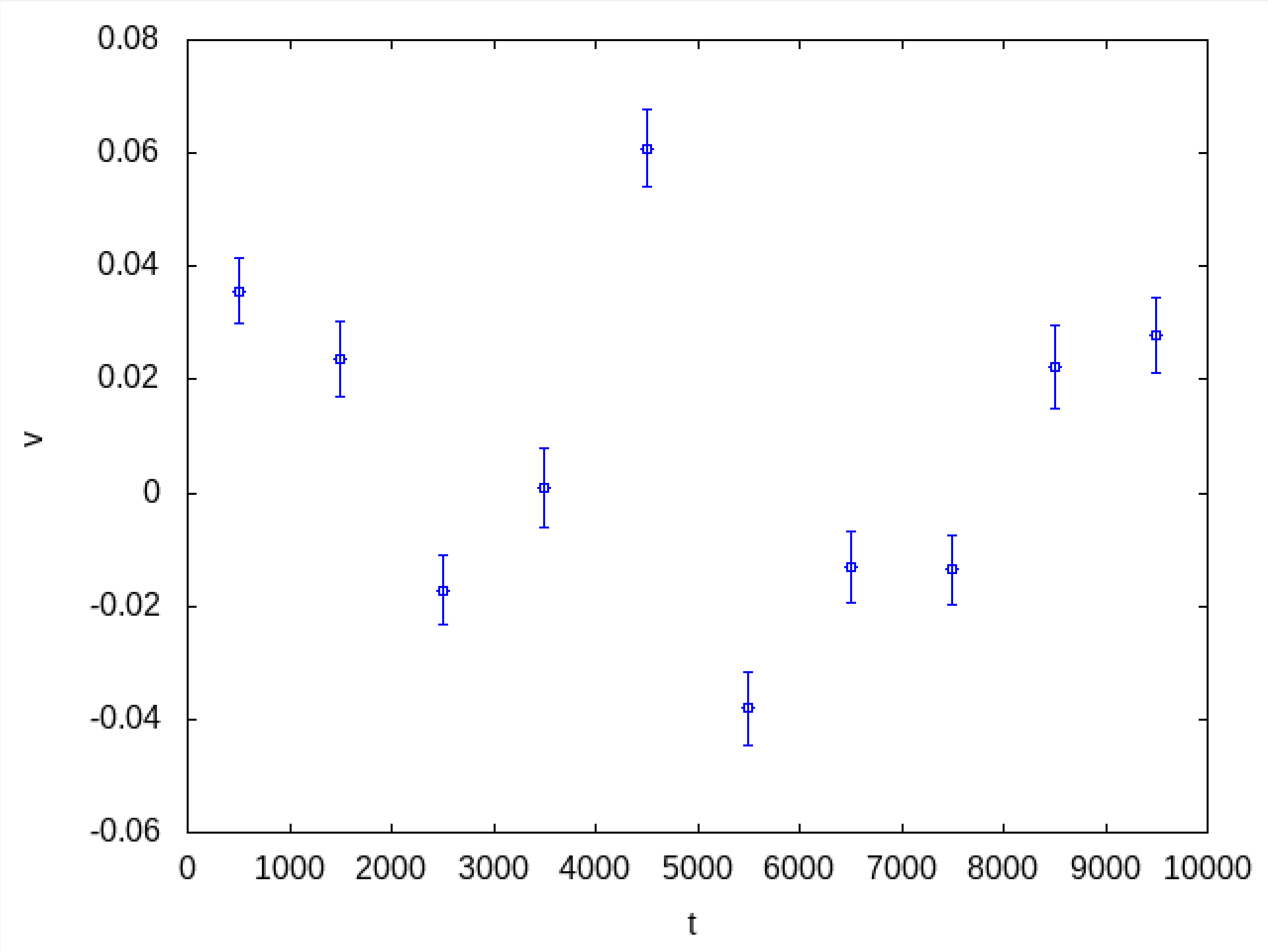
データ点ごとに異なる
データセットを生成している

適切なグラフ



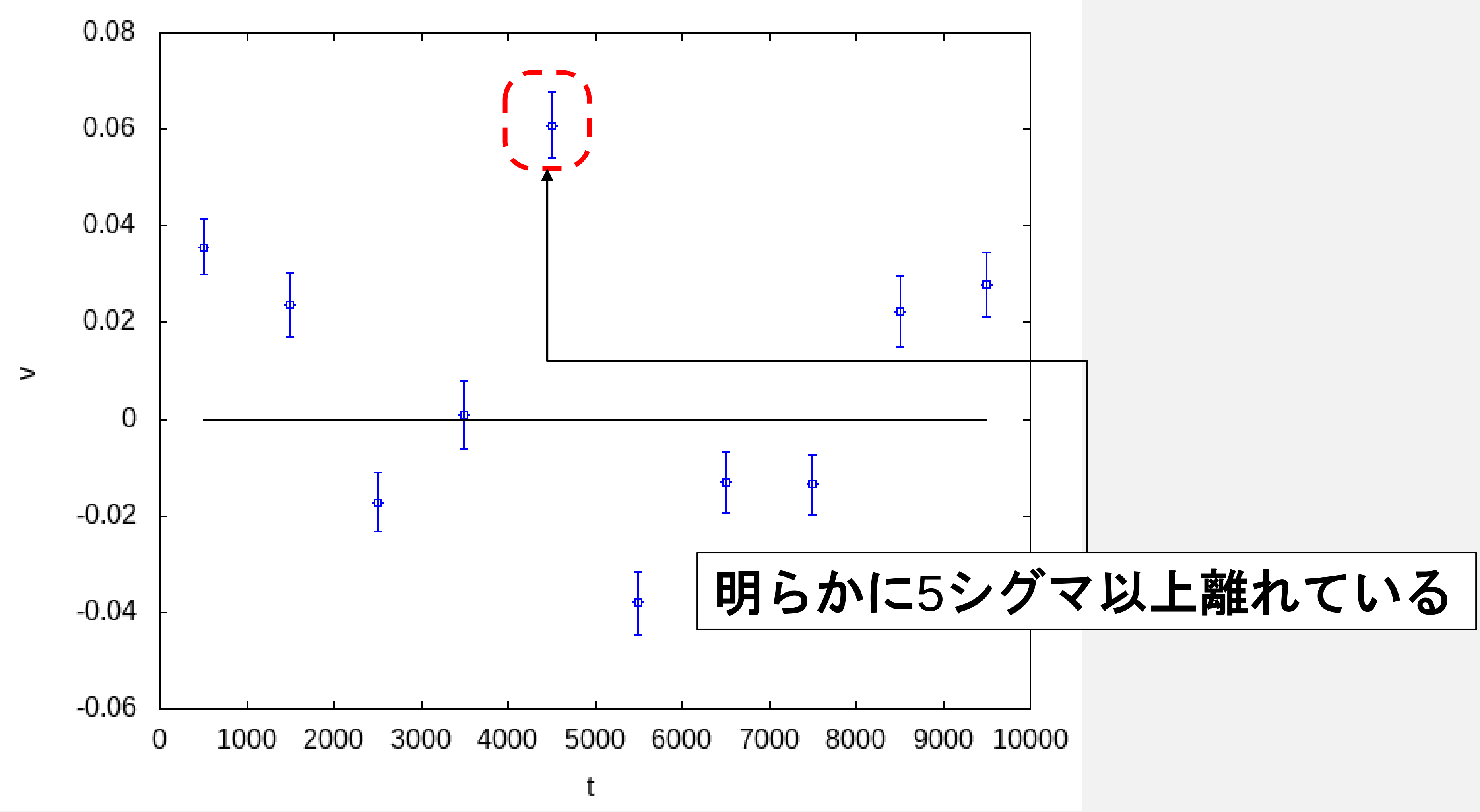
「真の値」の両側に均等にばらついており、もっともらしい

おかしなグラフ3



よほど複雑なデータでない限り、ゼロのまわりを揺らぐデータに見えるが・・・？

おかしなグラフ3


0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
v
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
t
明らかに5シグマ以上離れている

おかしなグラフ3

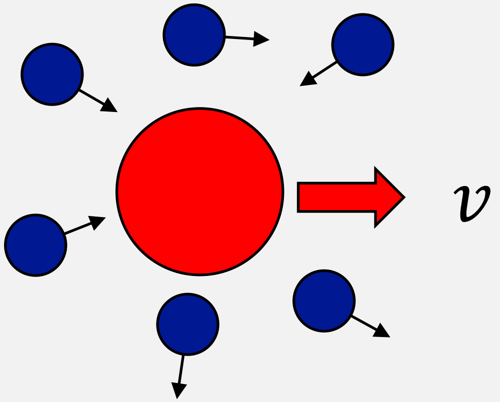
ランジュバン方程式の数値解法

```
import numpy as np

N = 1000
v = 0.0
gamma = 0.1
np.random.seed(1)

for j in range(10):
    d = np.zeros(N)
    for i in range(N):
        v += np.random.randn()*0.1
        v -= gamma * v
        d[i] = v
    ave = np.average(d)
    err = np.std(d) / np.sqrt(N)
    print(f"{(j+0.5)*N} {ave} {err}")
```

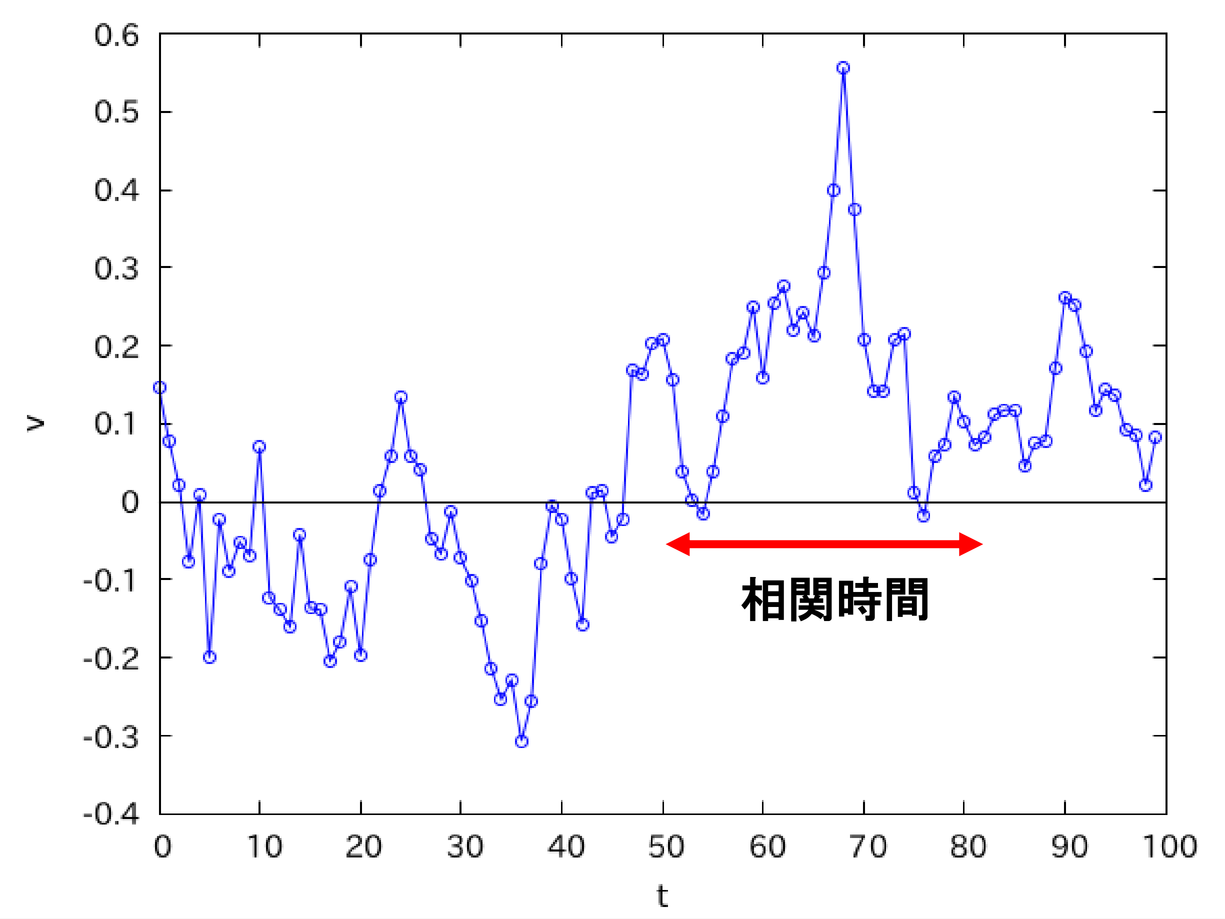
$$m\frac{dv}{dt} = -\gamma v + \hat{R}$$



1000ステップごとに平均、標準偏差を計算するコード

おかしなグラフ3

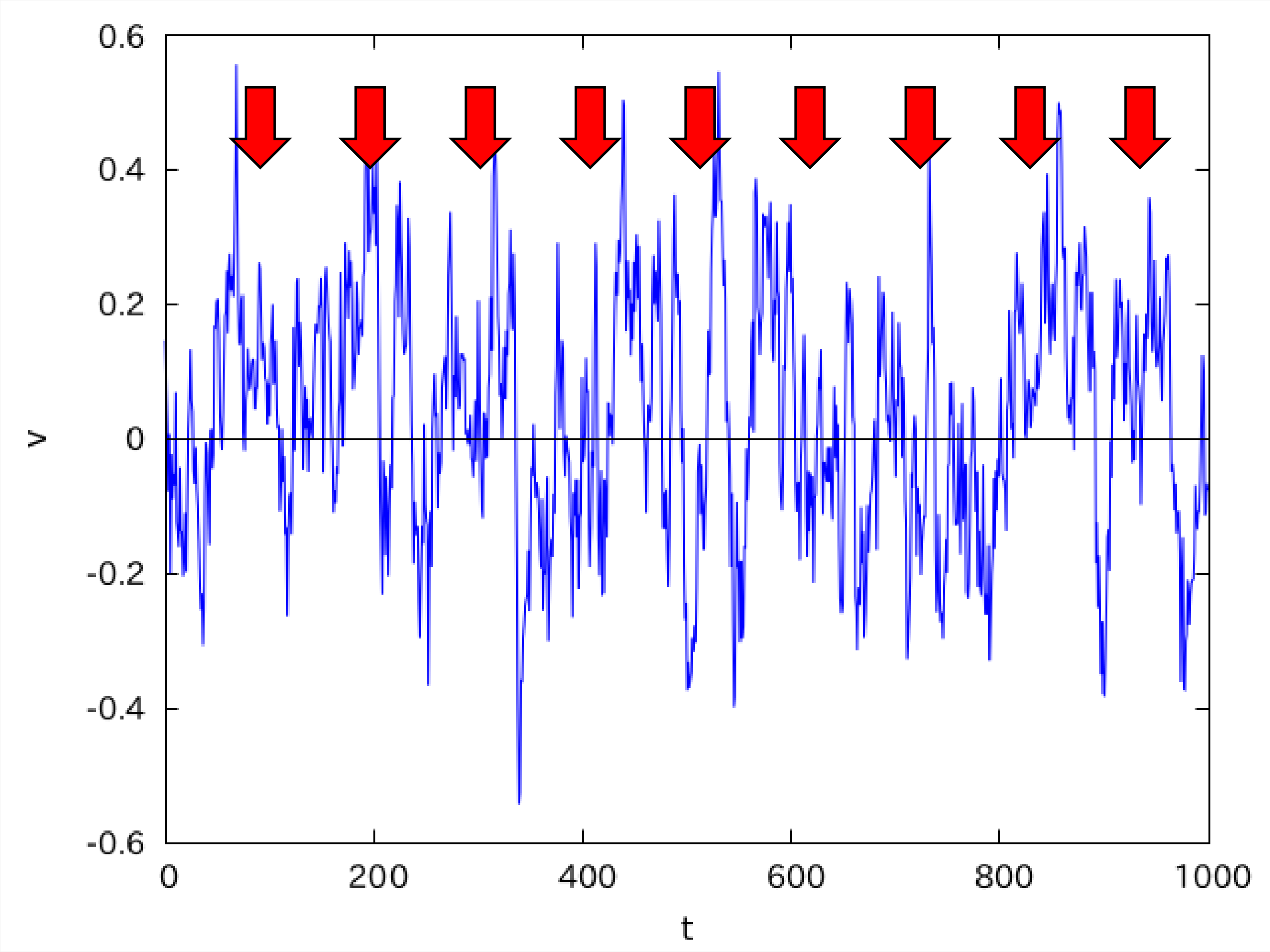
速度の時間発展データ



速度が「記憶」を失うまでにそれなりの時間がかかる

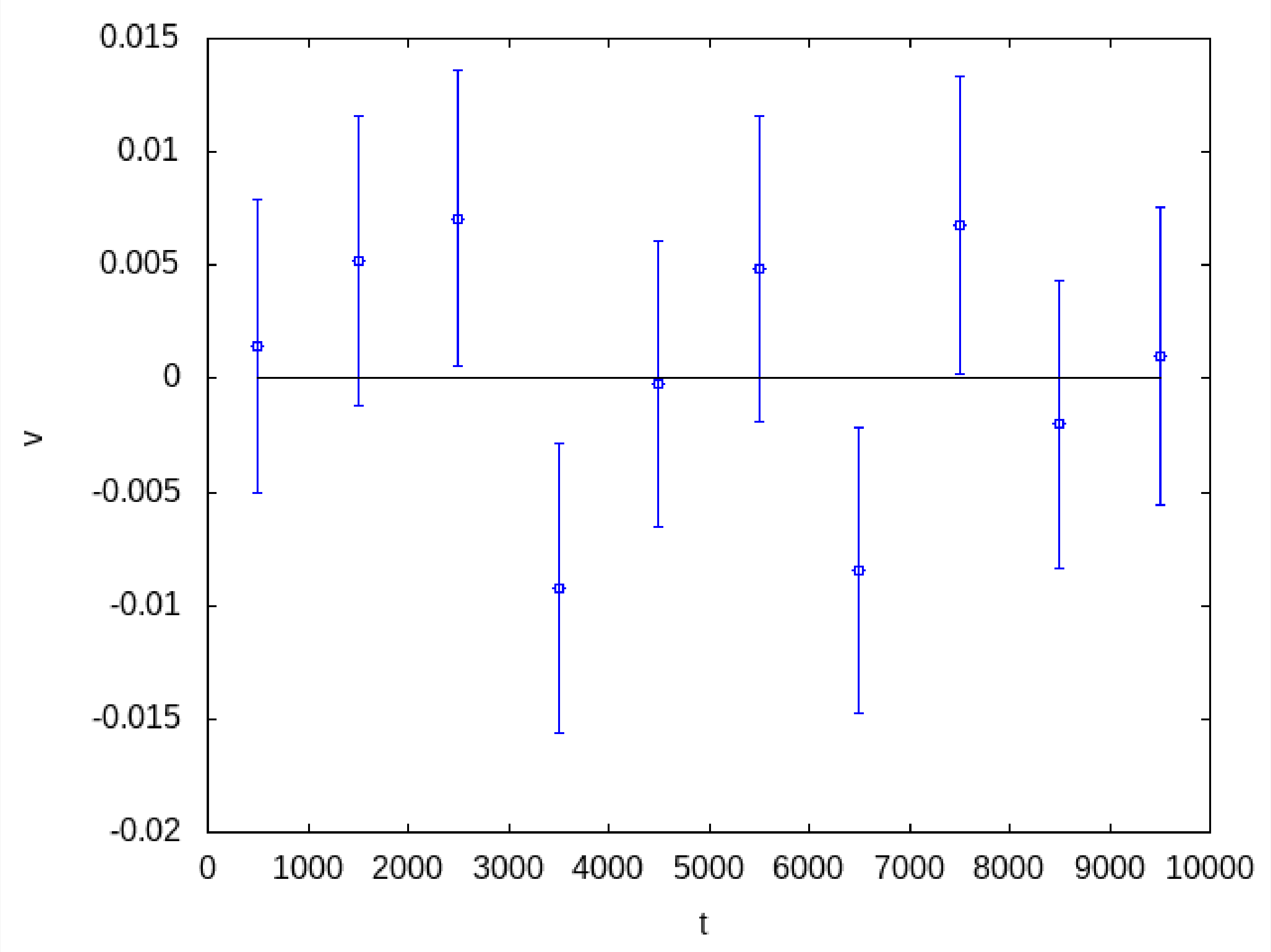
適切なグラフ

「記憶」を忘れそうな時間をあけて観測してサンプリングする



※もっとかしこい方法もある

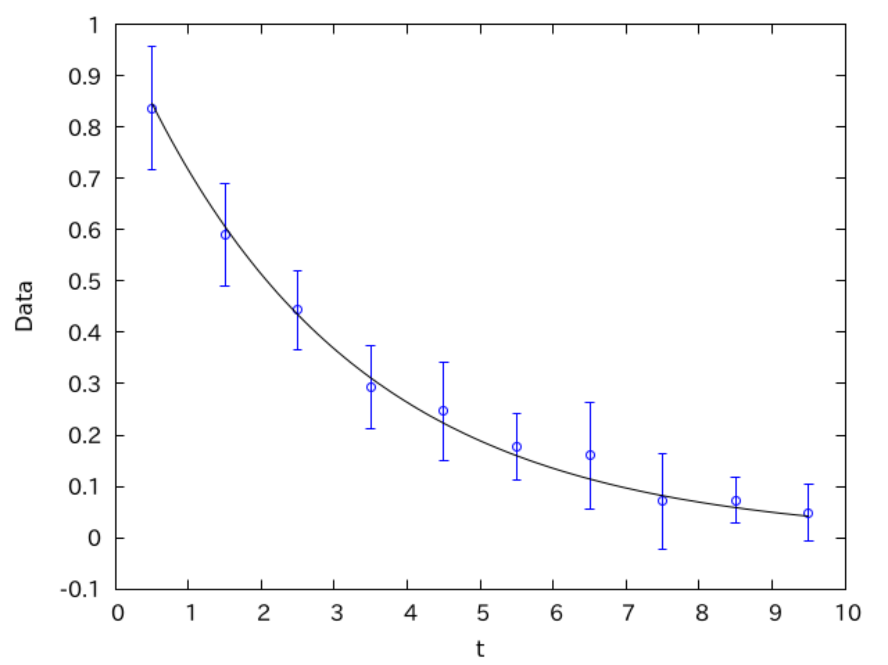
適切なグラフ



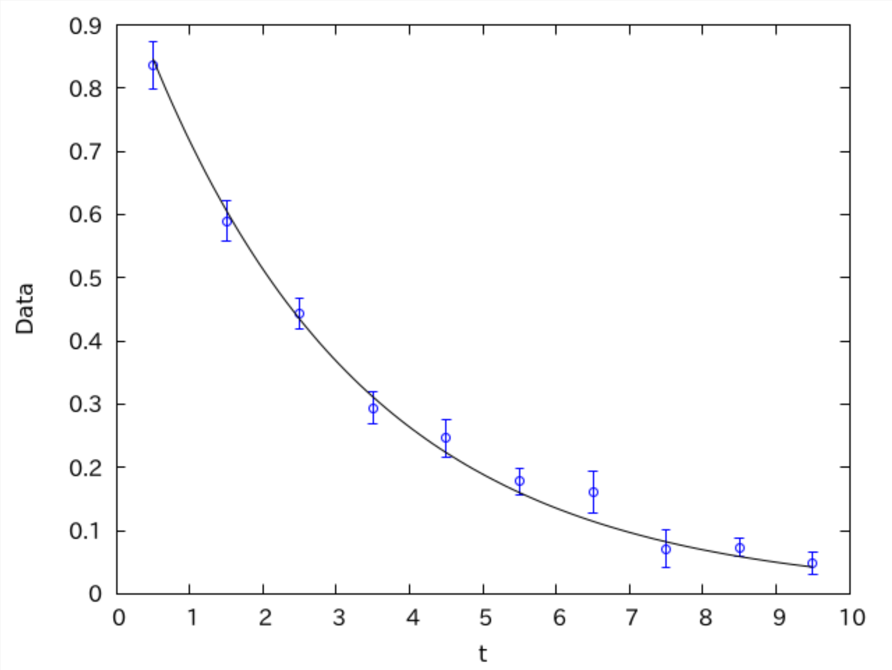
データのばらつき具合、エラーバーの外れ具合、ともにもっともらしい

不適切なグラフまとめ

エラーバーが大きすぎる



適切なグラフ

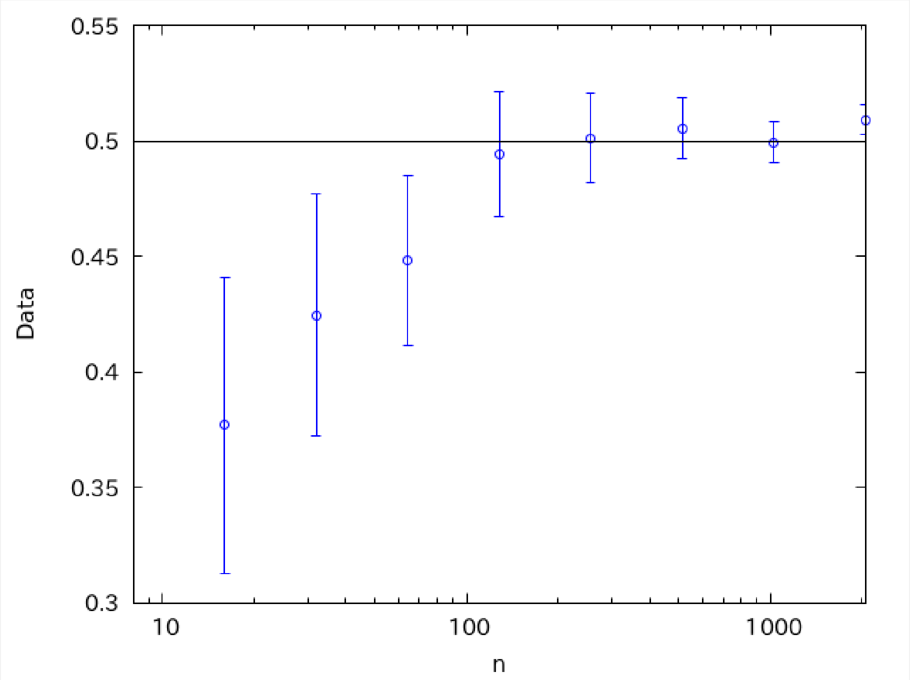


原因の例

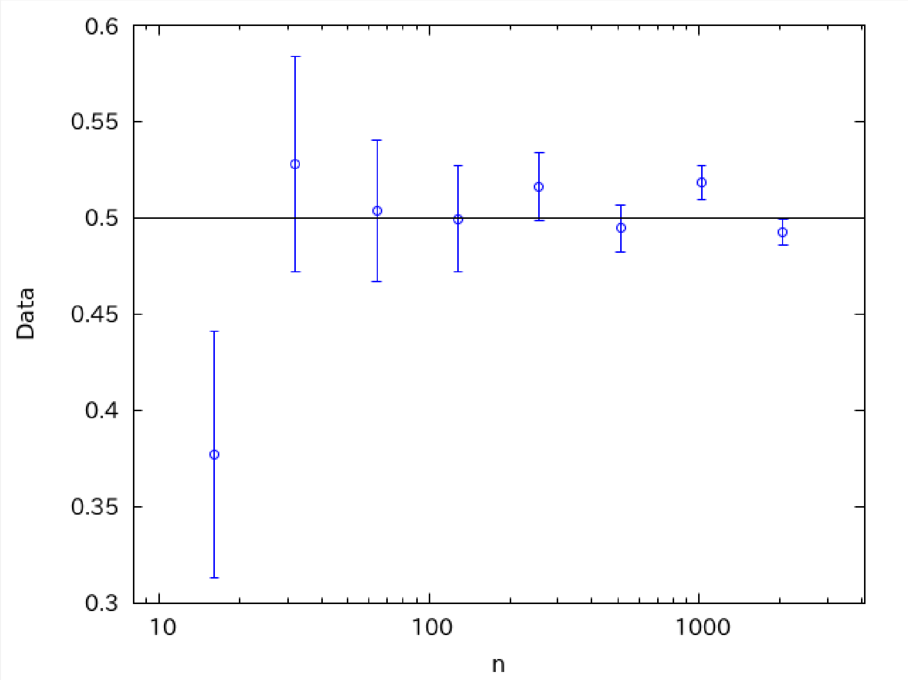
- $\sqrt{N}$で割り忘れている
- データに相関がある

不適切なグラフまとめ

偏りが大きい



適切なグラフ

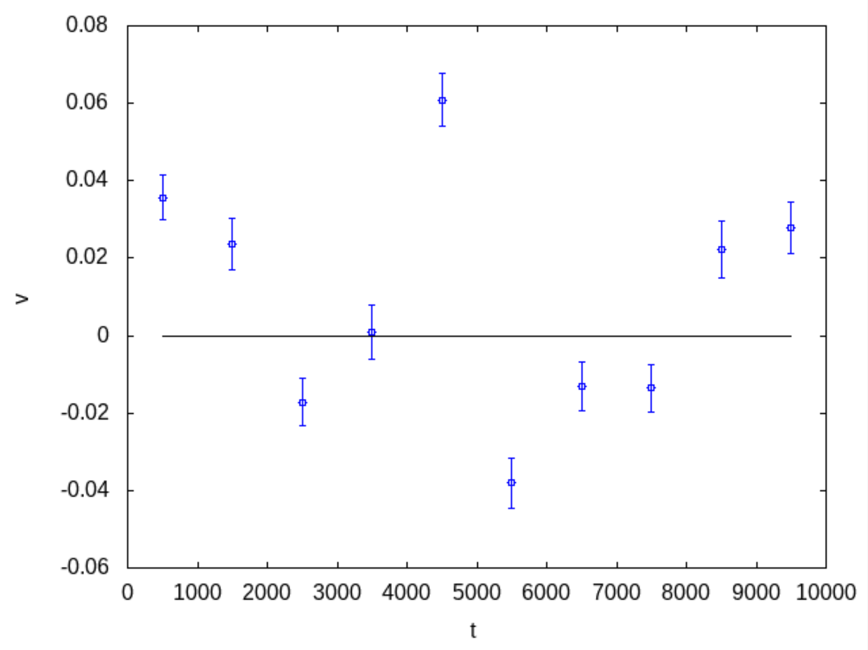


原因の例

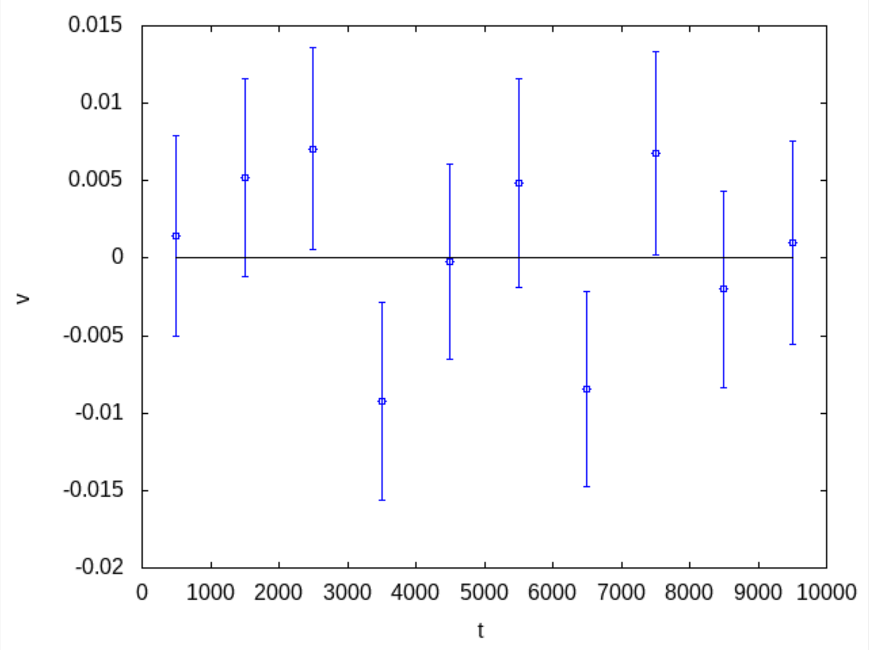
- データに相関がある

不適切なグラフまとめ

エラーバーが小さすぎる



適切なグラフ



原因の例

- データに相関がある

不適切なグラフのまとめ

データがガウス分布に従い、かつ独立であるなら

- 観測値は「真の値」の上下に均等にばらつく
- 観測値の3つに1つが「真の値」の1シグマの範囲に入らない
- 観測値と「真の値」がエラーバーの2倍離れることは稀、5倍離れることはまずない

逆に

- 観測値の全てが「真の値」をエラーバーの範囲に含む
- 「真の値」の片側に連続してずれている
- 「真の値」と5シグマ以上離れている

であるなら、何かがおかしい

エラーバーがおかしいグラフは、データの相関が原因であることが多い

不偏推定量とJackknife法

M個のデータがある

M

それをN個ずつのブロックに分割する

N

それぞれのブロックで期待値を計算する

μ_1 μ_2 $\mu_{M/N}$

期待値の期待値を計算する $\langle \mu \rangle = \dfrac{N}{M} \sum_i \mu_i$

ブロックサイズNを変えた時$\langle \mu \rangle$は変わるか？

不偏推定量とJackknife法

ブロックごとの期待値

$$\mu_i = \frac{1}{N}\sum_{k\in i} X_k$$

期待値の期待値

$$\langle\mu\rangle = \frac{N}{M}\sum_{i} \mu_i$$

単なる全体の平均になる

$$\langle\mu\rangle = \frac{1}{M}\sum_{j} X_j$$

$\langle\mu\rangle$はN依存性をもたない

不偏推定量とJackknife法

それぞれのブロックで期待値を計算する

μ_1	μ_2	$\cdots$							

それぞれのブロックの期待値の逆数を計算する

$1/\mu_1$	$1/\mu_2$	$\cdots$							

期待値の逆数の期待値を計算する $\langle 1/\mu \rangle = \frac{N}{M} \sum_i \frac{1}{\mu_i}$

$\langle 1/\mu \rangle$はN依存性を持つか？

不偏推定量とJackknife法

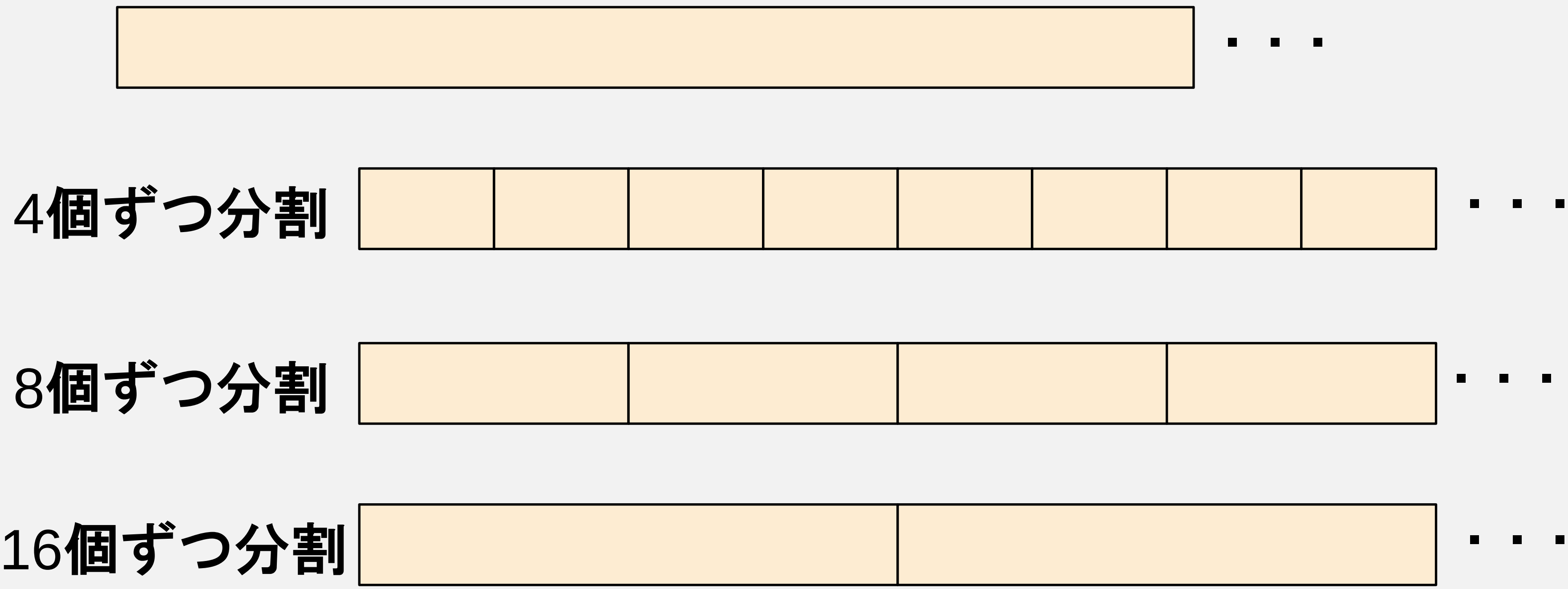
サイコロの目の期待値の逆数は？

期待値 $$\mu = \frac{1}{6}\sum_{k=1}^{6} k = \frac{7}{2} = 3.5$$

期待値の逆数 $$\frac{1}{\mu} = \frac{2}{7} \sim 0.286$$

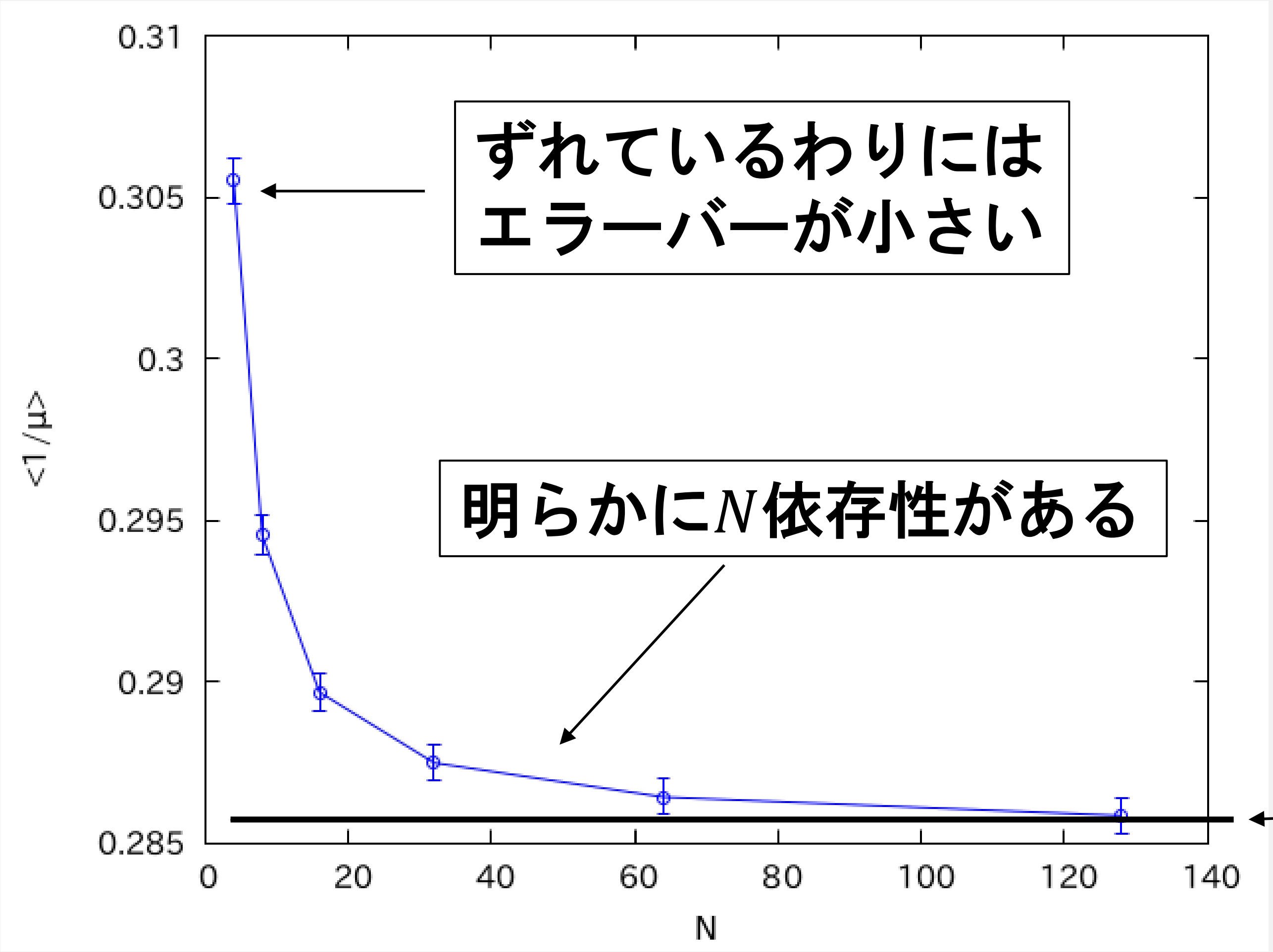
不偏推定量とJackknife法

サイコロを65536回振る



各ブロックで期待値μ_iを計算し、その逆数の期待値を計算する

不偏推定量とJackknife法



厳密解

同じデータセットを使っているのに、ブロックサイズが小さいところで挙動がおかしい→系統誤差

統計誤差と系統誤差

誤差(真値からのずれ)には、統計誤差と系統誤差の二種類がある

統計誤差 (statistical error)

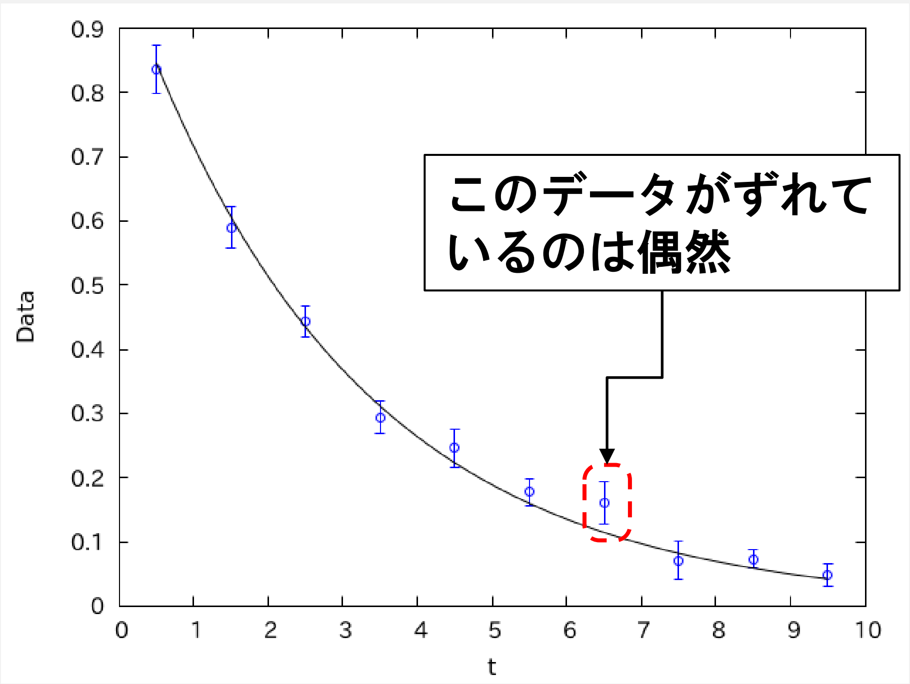
- 我々が制御できない要因により値が揺らぐこと(偶然誤差)
- 数値計算では乱数や粗視化に起因
- 不確かさ(uncertainty)とも

系統誤差 (bias, systematic error)

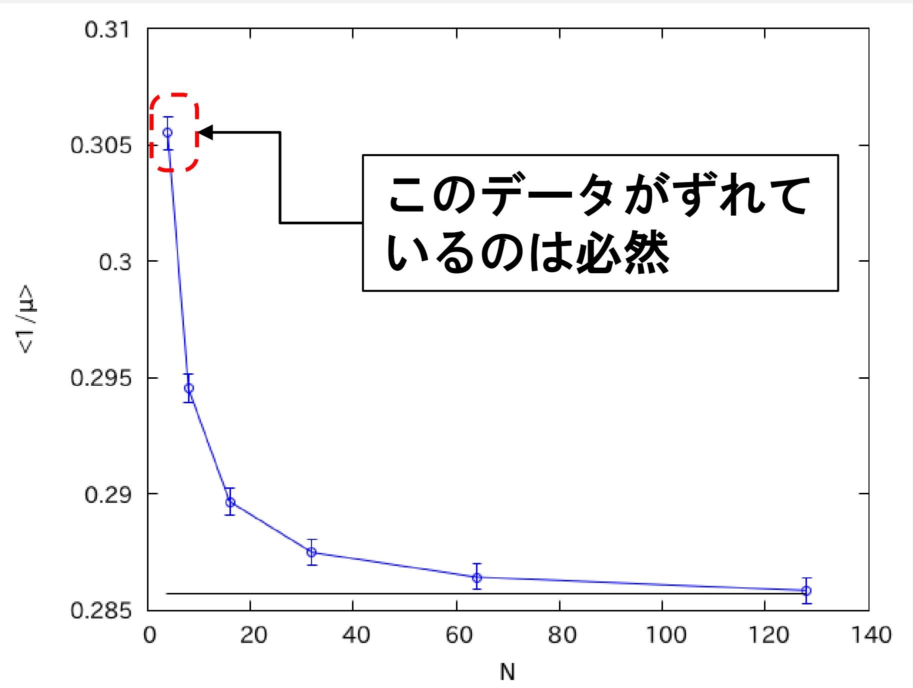
- 誤差を生む要因が説明できるもの
- 決定論的なずれ
- 数値計算では有限サイズ効果や理論誤差などに起因

統計誤差と系統誤差

統計誤差



系統誤差



- この系統誤差はどこからくるのか？
- どうやって減らすか

を知るのがこの節の目的

期待値の関数

確率変数$\hat{X}$の期待値μの関数の値を推定したい

$$y = g(\mu)$$

N回測定して期待値を推定する(これも確率変数)

$$\hat{\mu}_N = \frac{1}{N}\sum_i X_i$$

推定値の期待値は期待値に一致する

$$\langle \hat{\mu}_N \rangle = \mu$$

推定値の関数の期待値は期待値の関数と一致しない

$$\langle g(\hat{\mu}_N) \rangle \neq g(\mu)$$

不偏推定量

標本から得られた推定量(estimator)の期待値が母集団の期待値と一致する時、その推定量を不偏推定量(unbiased estimator)と呼ぶ

例：確率変数$\hat{X}$のN個のサンプル$\{X_i\}$から母集団$\{\hat{X}\}$の期待値μと分散σ^2を求めたい

期待値の推定値を
使っているのがポイント

$$\hat{\mu}_N = \frac{1}{N}\sum_i^N X_i$$

$$\hat{\sigma}_N^2 = \frac{1}{N}\sum_i^N (X_i - \hat{\mu}_N)$$

$$\langle \hat{\mu}_N \rangle = \mu$$

$$\langle \hat{\sigma}_N^2 \rangle = \frac{N-1}{N}\sigma^2 \neq \sigma^2$$

期待値は一致する(不偏推定量)

分散は一致しない(不偏推定量ではない)

期待値の関数

一般に確率変数 $\hat{X}$ について

関数の期待値 $\langle g(\hat{X}) \rangle$ と

期待値の関数 $g\left(\langle \hat{X} \rangle\right)$ は

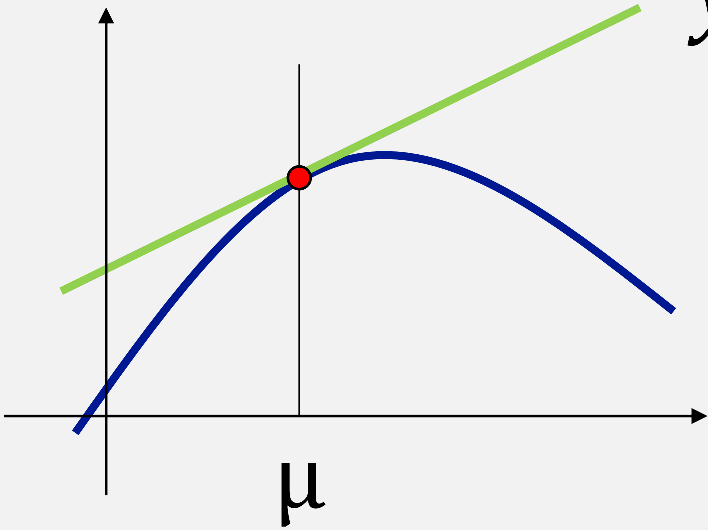
$$\left\langle g(\hat{X}) \right\rangle \neq g\left(\langle \hat{X} \rangle\right)$$ 一致しない

期待値の関数は、期待値の関数の不偏推定量ではない

Jensenの不等式

$g(x)$を上に凸な関数とし、$x = \mu$で接線をひく

$$y = a(x - \mu) + g(\mu)$$



$$y = g(x)$$

上図より明らかに $g(x) \leq a(x - \mu) + g(\mu)$

両辺の期待値を取れば $\langle g(x) \rangle \leq g(\mu) = g(\langle x \rangle)$

下に凸の場合は符号が逆になる

※ Thanks to @genkuroki

期待値の関数

$$\hat{\mu}_N = \frac{1}{N}\sum_i \hat{X}_i$$ N回の測定で得られた期待値の推定量

$$\varepsilon = \hat{\mu}_N - \mu$$ 真の期待値とのずれ

$$g(\hat{\mu}_N) - g(\mu) = g(\mu + \varepsilon) - g(\mu)$$

推定値　　真の値

$$= g'(\mu)\varepsilon + \frac{1}{2}g''(\mu)\varepsilon^2 + O(\varepsilon^3)$$

$$\langle g(\hat{\mu}_N) - g(\mu)\rangle \sim \frac{1}{2}g''(\mu)\langle\varepsilon^2\rangle = \frac{g''(\mu)\sigma^2}{2N}$$

推定値と真の値のずれの期待値

期待値の推定値の分散

N依存性

期待値の関数

N個　のサンプルから推定した期待値の関数と、
真の期待値の関数のずれは$1/N$に比例する

$$\langle g(\hat{\mu}_N) - g(\mu) \rangle \propto \frac{1}{N}$$

これを1/Nバイアスと呼ぶ

関数$g(x)$の二階微分がゼロ(線形)である場合はバイアスは生じない

$$\langle g(\hat{\mu}_N) - g(\mu) \rangle \sim \frac{g''(\mu)\sigma^2}{2N}$$

特に$g(x) = x$の場合

$$\langle \hat{\mu}_N \rangle = \mu$$

1/Nバイアスの具体例

平均0、分散σ^2のガウス分布に従う確率変数Xを考える

$$\langle \hat{X}^2 \rangle = \sigma^2$$ 2次のモーメント

$$\langle \hat{X}^4 \rangle = 3\sigma^4$$ 4次のモーメント

4次と2次のモーメントの比を取ると、分散依存性が消える

$$\frac{\langle \hat{X}^4 \rangle}{\langle \hat{X}^2 \rangle^2} = 3$$ 尖度(Kurtosis)

この量の1/Nバイアスを確認する

1/Nバイアスの具体例

N個のサンプリング(N回の測定)で得られたデータから2次と4次のモーメントを推定する

$$\left\langle \hat{X}^2 \right\rangle_N = \frac{1}{N}\sum_i \hat{X}_i^2 \quad \left\langle \hat{X}^4 \right\rangle_N = \frac{1}{N}\sum_i \hat{X}_i^4$$

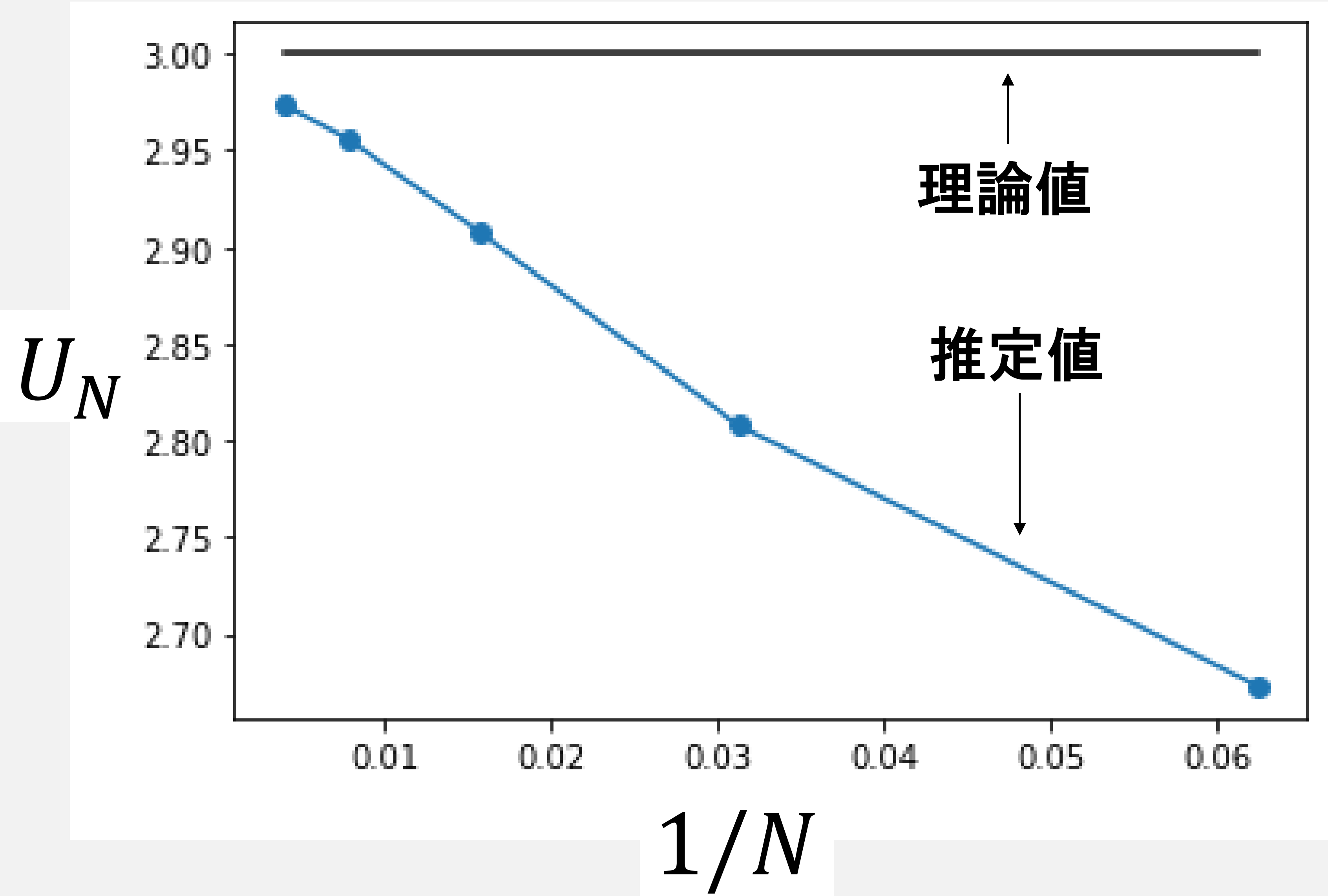
得られたモーメントから尖度を計算する

$$\hat{U}_N = \frac{\left\langle \hat{X}^4 \right\rangle_N}{\left\langle \hat{X}^2 \right\rangle_N^2}$$

上記を十分に繰り返して$\hat{U}_N$の期待値$\left\langle \hat{U}_N \right\rangle$を計算する

→ 統計誤差を消し、系統誤差だけを残す

1/Nバイアスの具体例



十分なサンプリング回数にも関わらず、真の値からずれている(バイアス)

統計誤差と系統誤差

不偏推定量ではあるが、ばらつきのせいで真の値からずれる誤差を**統計誤差**と呼ぶ

$$\hat{\mu}_N = \frac{1}{N}\sum_i \hat{X}_i \qquad \hat{\mu}_N - \mu = \boxed{O(1/\sqrt{N})}$$

不偏推定量でない推定量の期待値について、真の値からのずれを**系統誤差(バイアス)**と呼ぶ。

$$\langle g(\hat{\mu}_N)\rangle - g(\mu) = \boxed{O(1/N)}$$

サンプル数を増やすと統計誤差は減るが、
系統誤差は減らせない

バイアスの除去

期待値の関数の推定には1/Nバイアスが乗る
N無限大極限では一致するが、収束が遅い
手持ちのデータから1/Nバイアスを除去したい

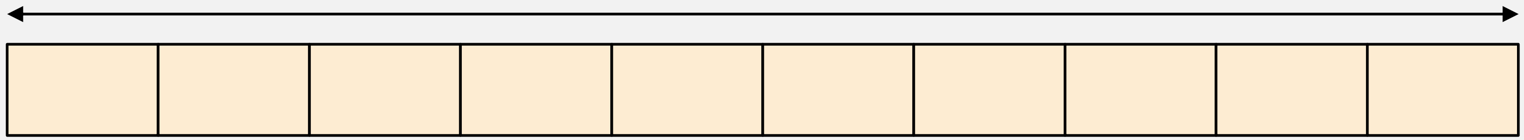
⬇

Jackknifeリサンプリング

バイアス除去

N個のデータがある

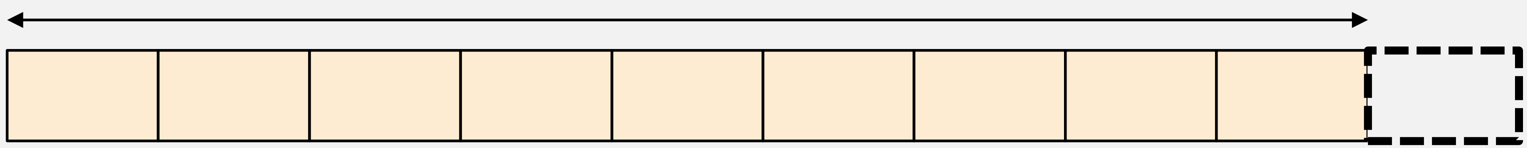
N



全部のデータを使って期待値μ_Nを計算
それを使って関数の推定値$U_N = g(\mu_N)$を計算

1個のデータを捨てる

$N - 1$



残りのデータを使って期待値μ_{N-1}を計算
それを使って関数の推定値$U_{N-1} = g(\mu_{N-1})$を計算

バイアス除去

U_Nは、真の値U_∞に対して1/Nバイアスがあると仮定

$$U_N = U_\infty + a/N$$

一つデータを捨てて得たU_Nのバイアスは

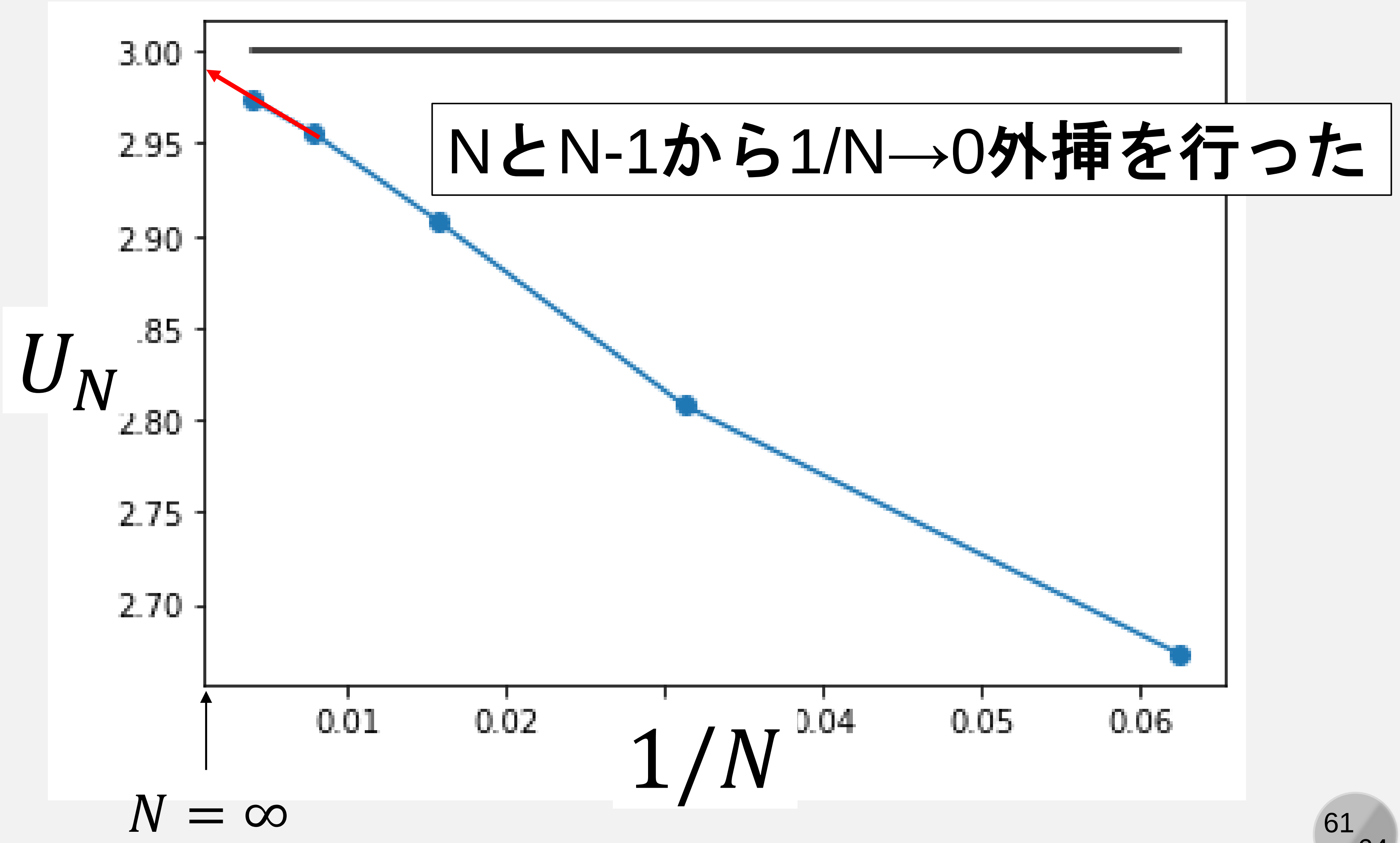
$$U_{N-1} = U_\infty + a/(N-1)$$

この2式からU_∞を求めると

$$U_\infty = NU_N - (N-1)U_{N-1}$$

※ Thanks to smorita and yomichi

バイアス除去

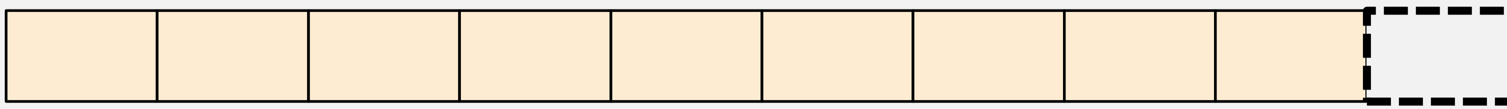


3.00
2.95
2.90
.85
2.80
2.75
2.70
0.01
0.02
0.04
0.05
0.06
U_N
$1/N$
$N = \infty$
NとN-1から1/N→0外挿を行った

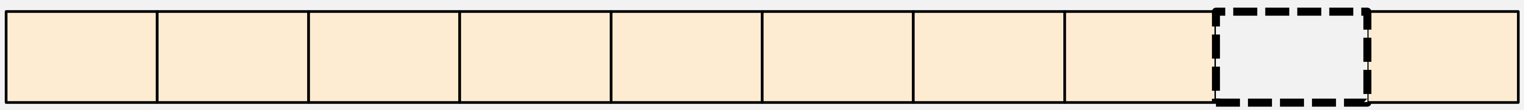
Jackknifeリサンプリング

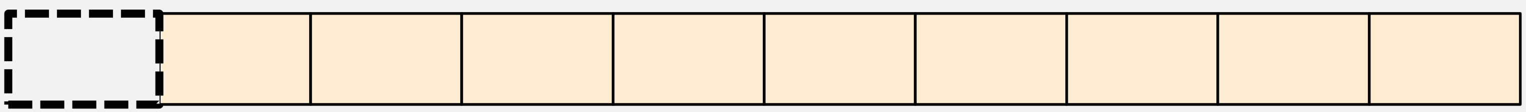
せっかくのデータを捨てるのはもったいないので活用する

1個のデータ除外して計算

U_{N-1}^{1} 

別のデータ除外して計算

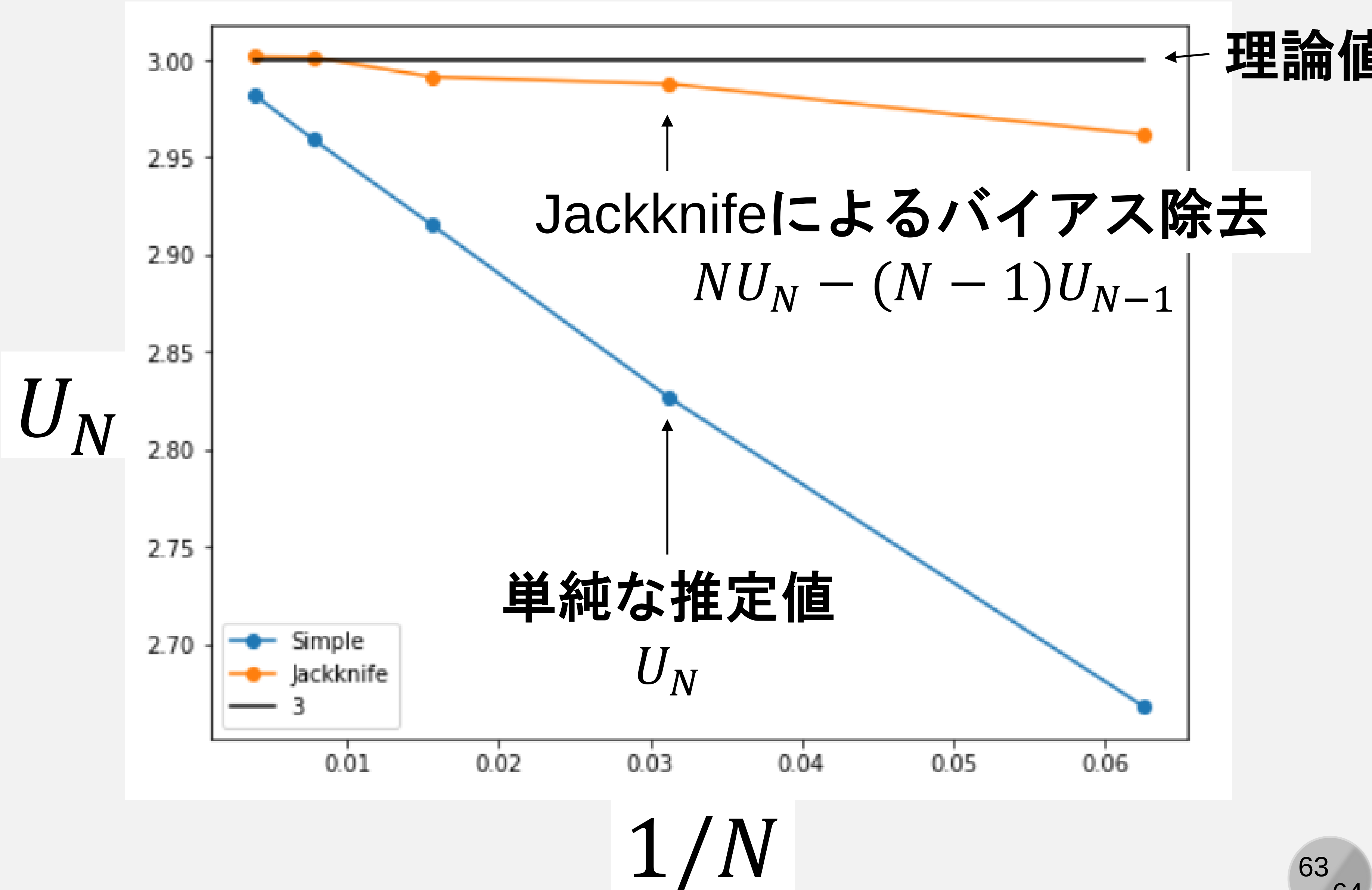
U_{N-1}^{2} 

U_{N-1}^{N} 

$$U_{N-1} = \frac{1}{N} \sum_i U_{N-1}^{i}$$

精度の高い「N-1個のデータの推定量」が得られる

Jackknifeリサンプリング



まとめ

- 母集団の何かを推定する量を推定量(estimator)と呼ぶ
- 誤差には統計誤差と系統誤差(バイアス)がある
- その期待値が母集団の期待値に一致する量(バイアスが無い量)を不偏推定量(unbiased estimator)と呼ぶ
- 期待値の関数の単純な推定は不偏推定量を与えない
- リサンプリングによりバイアスを除去できる
- Jackknife法はリサンプリング法の一種