# 保全生態学特論

## ― 統計的思考を身につける ―

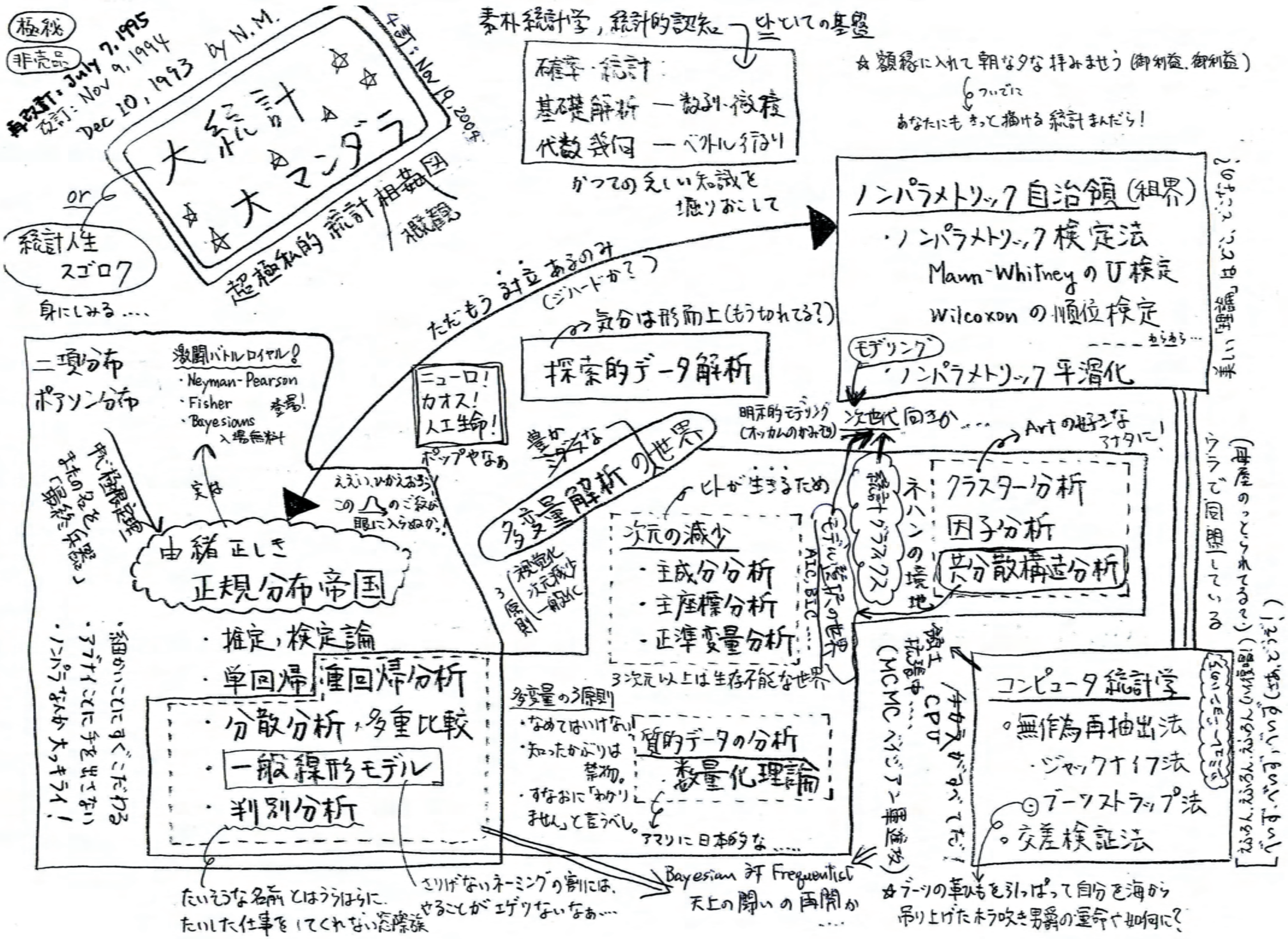
嘶役

三中 信宏

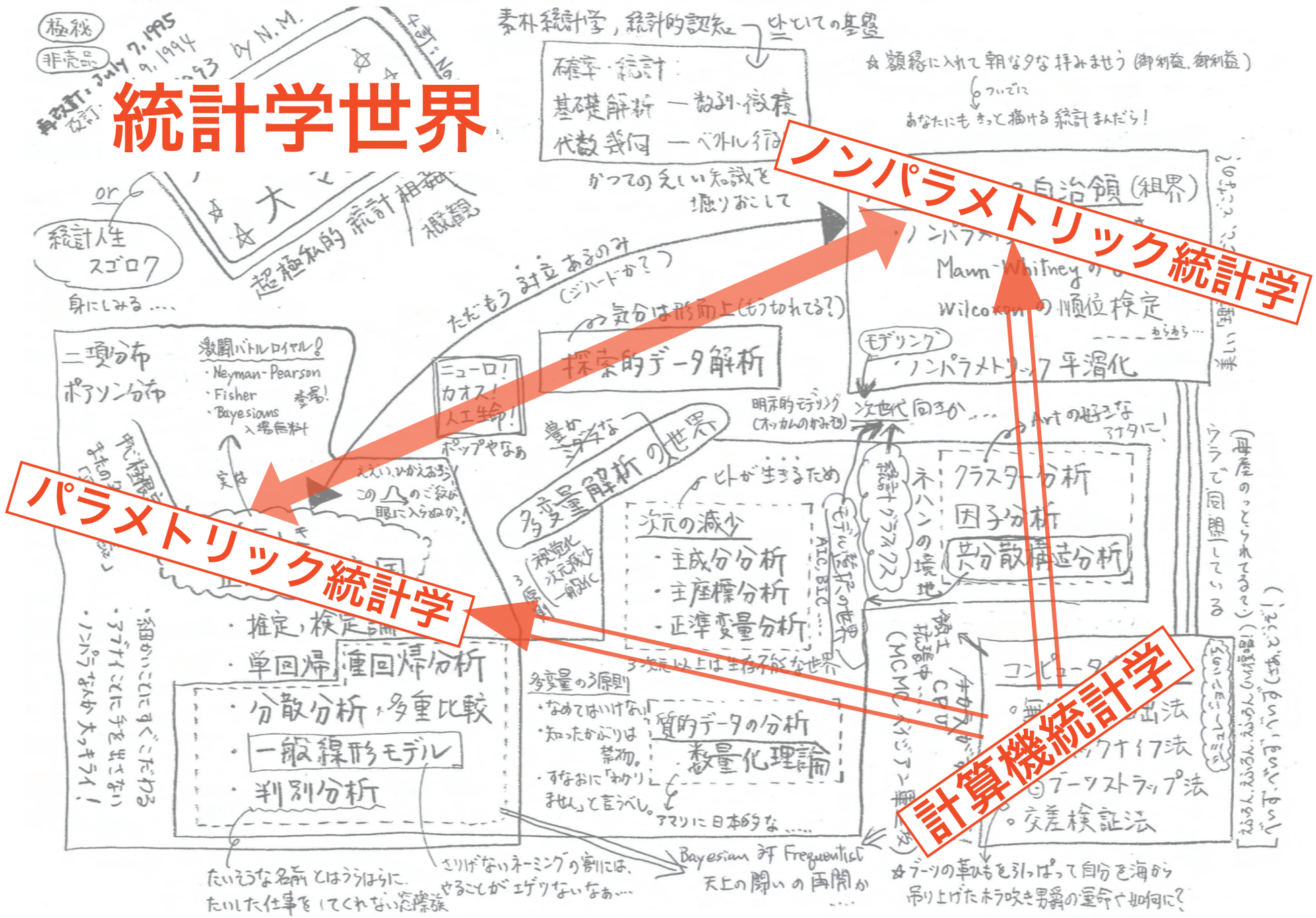
**MINAKA Nobuhiro**

国立研究開発法人 農研機構 農業環境変動研究センター ユニット長［生物統計学］
東京大学大学院 農学生命科学研究科 生物・環境工学専攻 教授［生態系計測学］
東京農業大学大学院 農学研究科 客員教授［応用昆虫学］

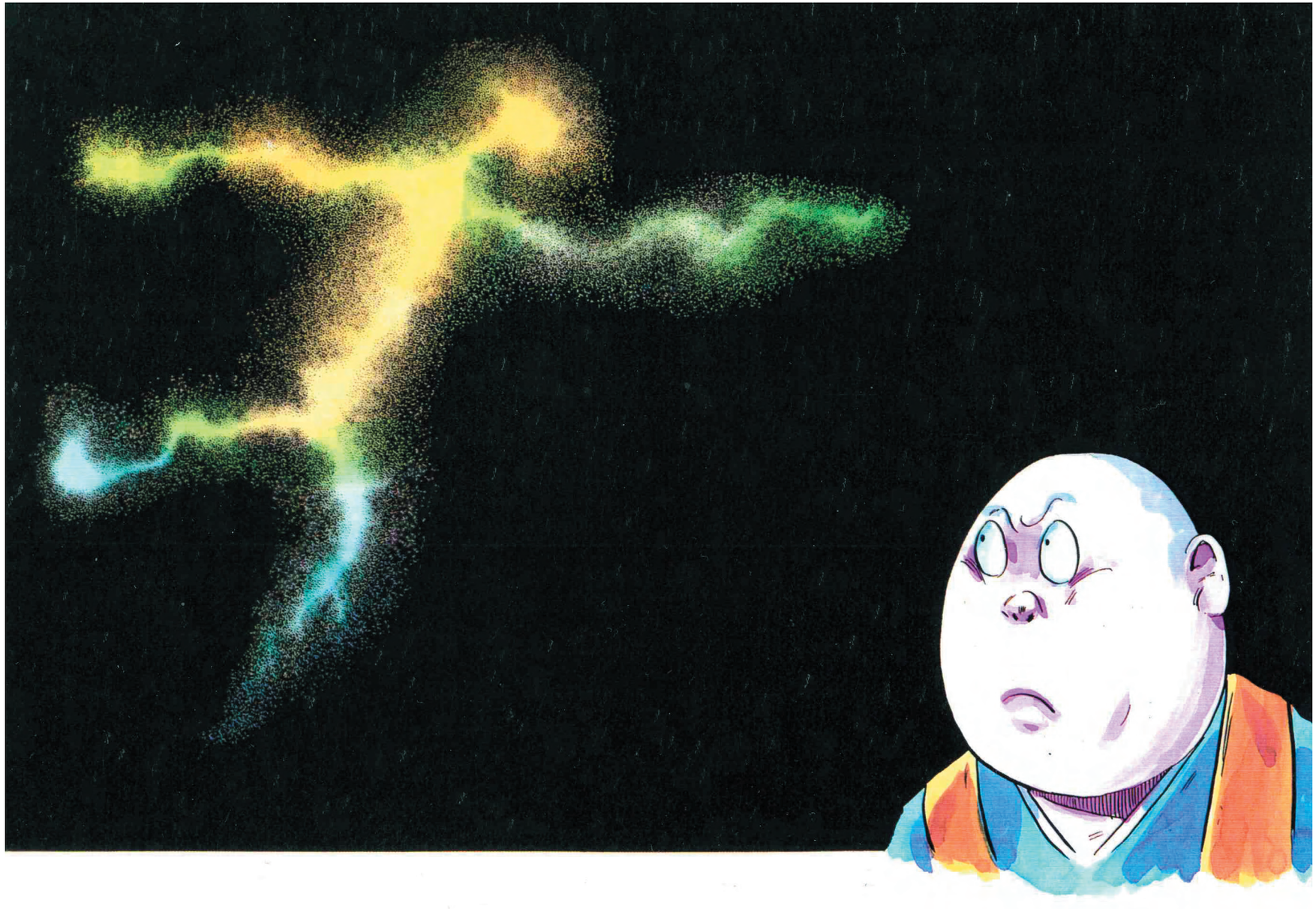
**mail: minaka@affrc.go.jp**
**website: http://leeswijzer.org/**
**twitter: @leeswijzer**



極秘
非売品
再改訂：July 7, 1995
改訂：Nov 9, 1994
Dec 10, 1993 by N. M.
大統計 大マンダラ
超極私的統計相姦図
概観
or
統計人生スゴロク
身にしみる….
素朴統計学，統計的認知 — 土台としての基盤
確率・統計
基礎解析 — 微分・積分
代数幾何 — ベクトル・行列
かつての えらい知識を 掘りおこして
☆ 額縁に入れて朝な夕な拝みませう（御利益，御利益）
ついでに
あなたにも きっと描ける 統計まんだら！
ノンパラメトリック自治領（租界）
・ノンパラメトリック検定法
Mann-Whitney の U 検定
Wilcoxon の順位検定
モデリング
・ノンパラメトリック平滑化
ただもう対立あるのみ（ジハードかて？）
気分は形而上（もうきれてる？）
探索的データ解析
ニューロ！ カオス！ 人工生命！
ポップやなぁ
明示的モデリング（オッカムのかみそり）
次世代向きか….
Artの好きなアナタに！
二項分布
ポアソン分布
激闘バトルロイヤル！
・Neyman-Pearson
・Fisher
・Bayesians 入場無料
登場！
実は
ええい，ひかえおろう！ この のご紋が眼に入らぬか！
由緒正しき 正規分布帝国
・推定，検定論
・単回帰 重回帰分析
・分散分析，多重比較
・一般線形モデル
・判別分析
豊かな多変量解析の世界
視覚化 次元縮少 一般化 3原則
ヒトが生きるため
次元の減少
・主成分分析
・主座標分析
・正準変量分析
AIC, BIC
モデル選択の世界
統計グラフィクス
ネハンの境地
クラスター分析
因子分析
共分散構造分析
3次元以上は生存不能な世界
多変量の3原則
・なめてはいけない
・知ったかぶりは禁物。
・すなおに「わかりません」と言うべし。
質的データの分析
数量化理論
アマリに日本的な….
ウラで同盟している
コンピュータ統計学
・無作為再抽出法
・ジャックナイフ法
・ブーツストラップ法
・交差検証法
CPU
MCMC ベイジアン再燃？
☆ブーツの革ひもを引っぱって自分を海から吊り上げたホラ吹き男爵の運命や如何に？
Bayesian 対 Frequentist
天上の闘いの再開か….
たいそうな名前とはうらはらに たいした仕事をしてくれない窓際族
さりげないネーミングの割には，やることがエゲツないなぁ….
・細かいことにすぐこだわる
・アブナイことに手を出さない
・ノンパラなんか大ッキライ！

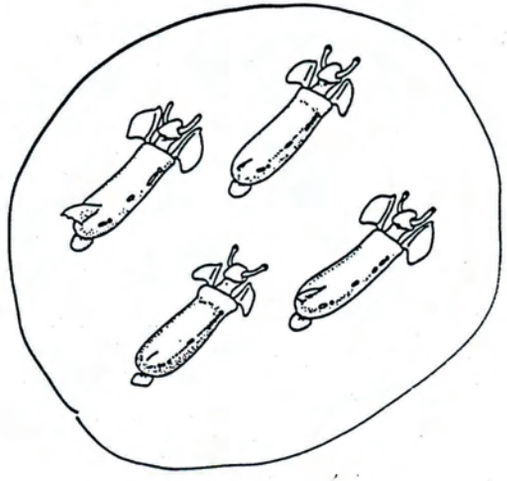


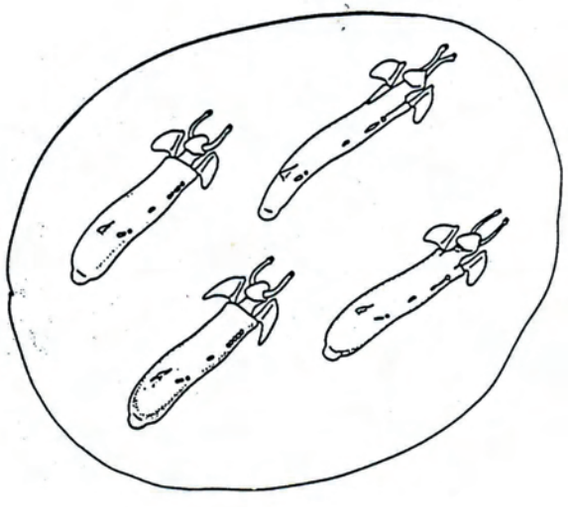

第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法



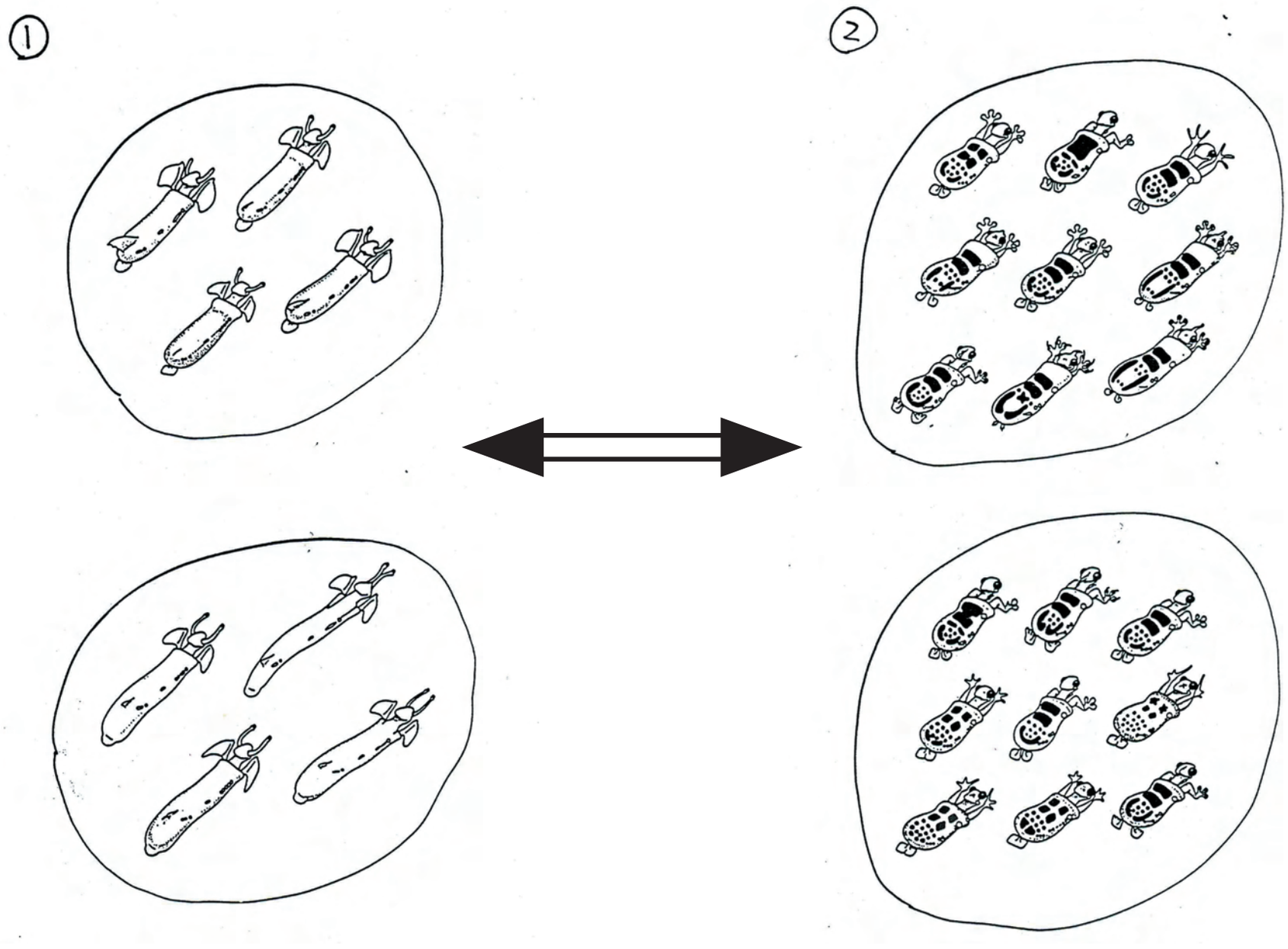
# 等身大の「科学」

①





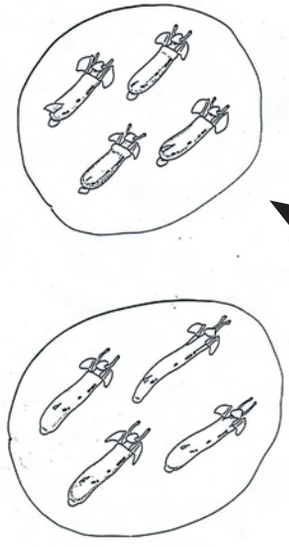
2


①
②

# ヒトの素朴統計学的「直感」

いつ「差がある」と感じるか？

①



○ケース１）

「群間のちがい／群内のちがい」が大きいとき

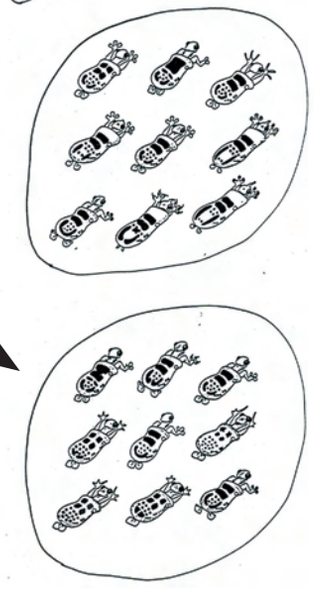
→「差がある」と感じる

○ケース２）

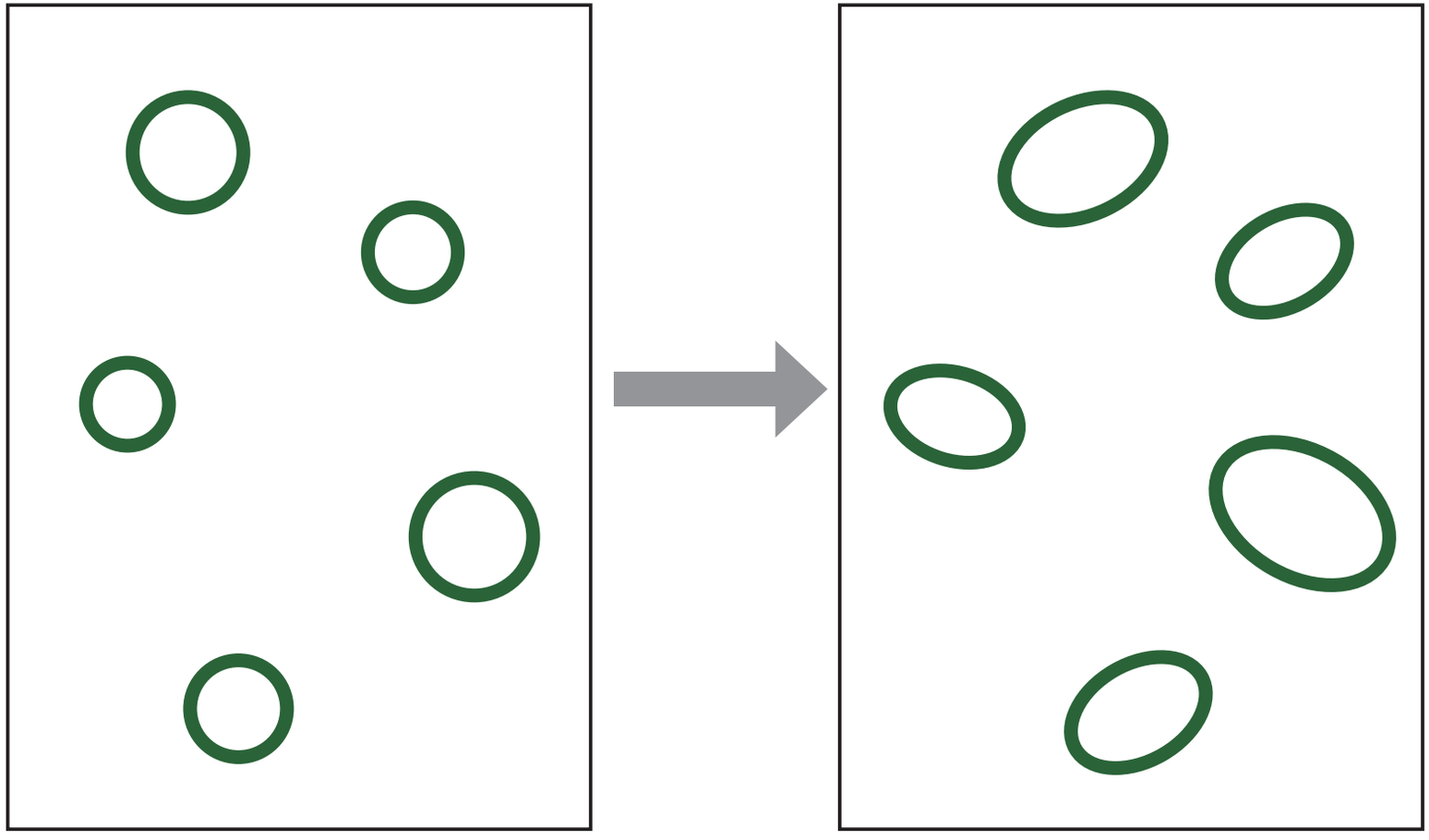
「群間のちがい／群内のちがい」が小さいとき

→「差がない」と感じる

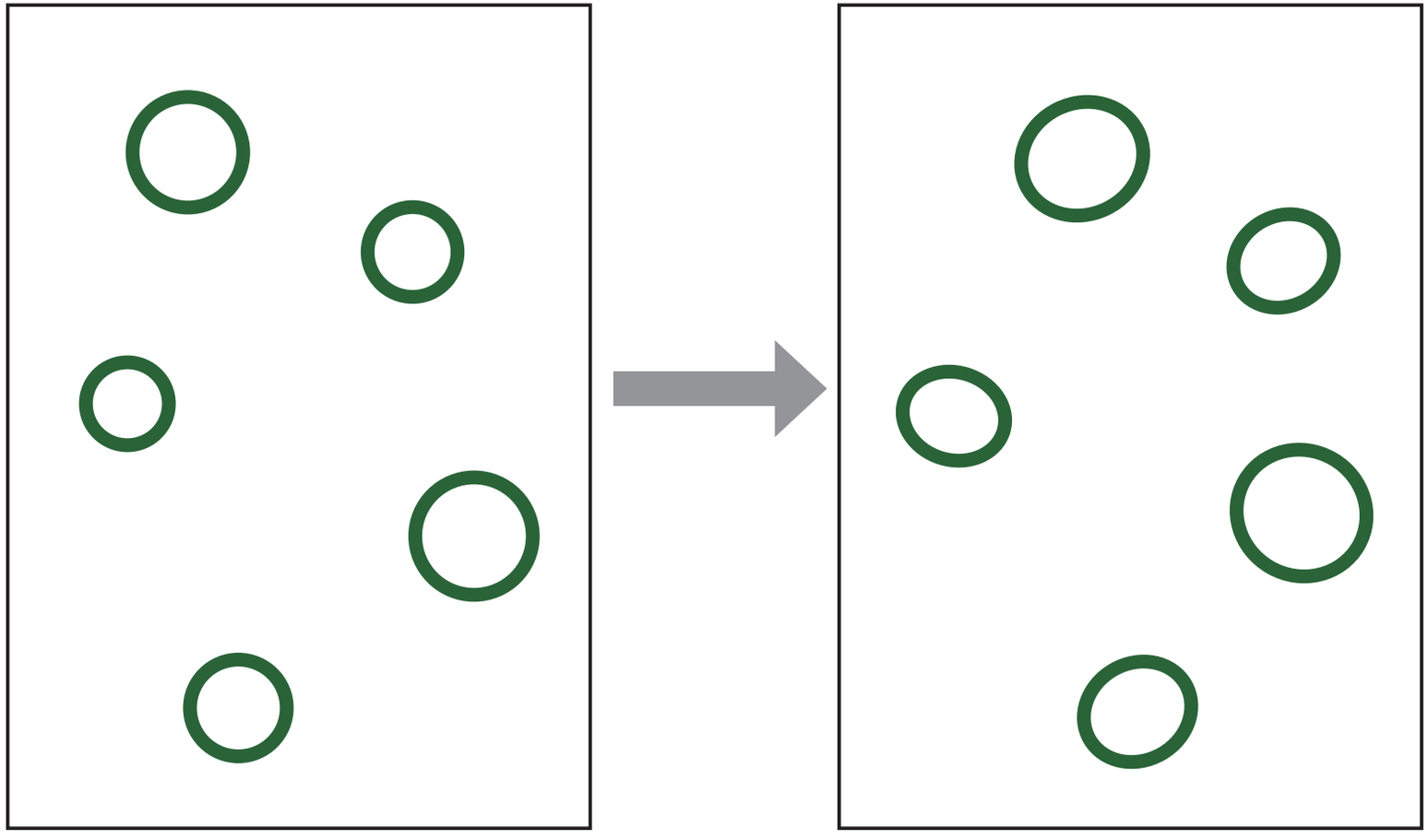
②



※判定基準となる比率＝「群間のちがい／群内のちがい」を定量化すると，統計学でいう「Ｆ値」になる．




【図 1】




【図 2】

ELSEVIER

Cognition 58 (1996) 1–73

COGNITION

# Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty

Leda Cosmides*, John Tooby

*Center for Evolutionary Psychology, University of California, Santa Barbara, CA 93106, USA*

Received September 1, 1992, final version accepted January 2, 1995

**Abstract**

Professional probabilists have long argued over what probability means, with, for example, Bayesians arguing that probabilities refer to subjective degrees of confidence and frequentists arguing that probabilities refer to the frequencies of events in the world. Recently, Gigerenzer and his colleagues have argued that these same distinctions are made by untutored subjects, and that, for many domains, the human mind represents probabilistic information as frequencies. We analyze several reasons why, from an ecological and evolutionary perspective, certain classes of problem-solving mechanisms in the human mind should be expected to represent probabilistic information as frequencies. Then, using a problem famous in the "heuristics and biases" literature for eliciting base rate neglect, we show that correct Bayesian reasoning can be elicited in 76% of subjects – indeed, 92% in the most ecologically valid condition – simply by expressing the problem in frequentist terms. This result adds to the growing body of literature showing that frequentist representations cause various cognitive biases to disappear, including overconfidence, the conjunction fallacy, and base-rate neglect. Taken together, these new findings indicate that the conclusion most common in the literature on judgment under uncertainty – that our inductive reasoning mechanisms do not embody a calculus of probability – will have to be re-examined. From an ecological and evolutionary perspective, humans may turn out to be good intuitive statisticians after all.

# 形而上学（＝「存在」の学）

「ある」のか？

「ない」のか？

# ないもの、

# あります

クラフト・エヴィング商會・著

Craft Ebbing & Co.

よく耳にはするけれど
一度として見たことのないものたち、
**あります。**

堪忍袋の緒、転ばぬ先の杖、左うちわ、舌鼓、
**あります。**

巻末には、赤瀬川原平氏書き下ろしエッセイ
『どりあえずビールでいいのか』
**あります。**

筑摩書房　定価（本体価格１４００円＋税）

# *Type I Error*

「ない」のに「ある」と言う過誤

**VS**

# *Type II Error*

「ある」のに「ない」と言う過誤

## 帰無仮説

処理効果（共通要因）は「ない」

VS

## 対立仮説

処理効果（共通要因）は「ある」

理髪師は急に声を低めた「この世界には〈解釈〉しか存在しておらんのじゃ．この世には，人の数だけ〈解釈〉があるのじゃ.」「真実などない，とそうおっしゃりたいのですか？」と探偵は鏡ごしに理髪師の目を見た。「私は探偵として，真実を擁護したい．探偵が真実を見失ったら，商売できなくなるのです。」

クラフト・エヴィング商會(1998)
『クラウド・コレクター：雲をつかむような話』
筑摩書房，東京，p.94

# 歴史・レトリック・立証

カルロ・ギンズブルグ

上村忠男訳

みすず書房

歴史・レトリック・立証

C・ギンズブルグ　上村忠男訳

みすず書房

歴史（ヒストリー）は虚構（フィクション）なのか。過去の出来事は復元できないのか。ニーチェからフローベールまで、表象と真実をめぐるギンズブルグの最新論集。

みすず書房　定価（本体2800円＋税）

表象と真実

みすず書房

資料は実証主義者たちが信じているように開かれた窓でもなければ、懐疑論者たちが主張するように視界をさまたげる壁でもない。いってみれば、それらは歪んだガラスにたとえることができるのだ。

カルロ・ギンズブルグ 2001
『歴史・レトリック・立証』
みすず書房 , p.48

ひとは証拠を逆撫でしながら、それをつくりだした者たちの意図にさからって、読むすべを学ばなければならない。

カルロ・ギンズブルグ 2001
『歴史・レトリック・立証』
みすず書房, p.48

# 推論の形式

**Deduction** （演繹）

前提から論理的に正しい結論を導く

**Induction** （帰納）

観察の蓄積により，一般則を導く

**Abduction** （仮説発見）

与えられた観察から，最良の仮説を発見する

# 「第三」の推論様式：

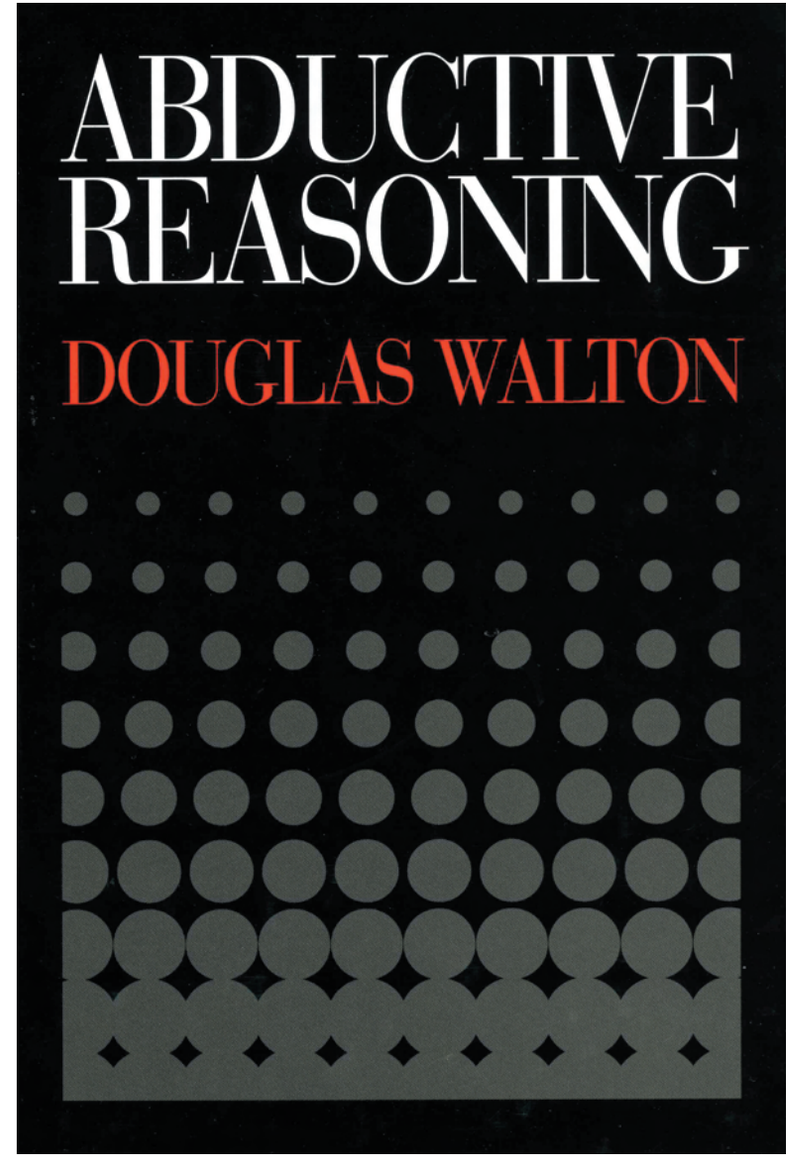
1) データ $D$ が得られた.

2) 仮説 $H$ は $D$ を説明できる

3) $H$ 以外の仮説 $H'$ は $H$ ほど うまく $D$ を説明できない.

4) したがって $H$ を受容する.

**Inference to the best explanation**



# 「ベスト」の説明？

# 最節約基準

# 尤度基準

# ベイズ基準

● ● ● ● ● ●

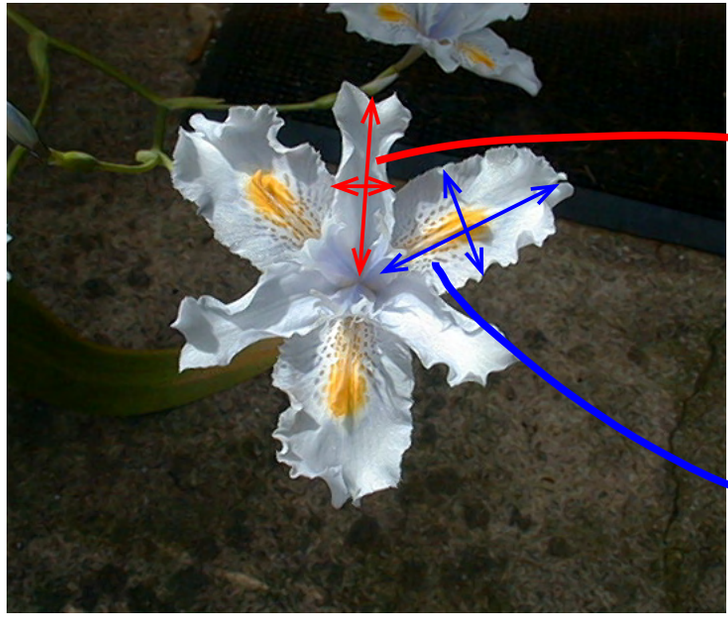
# データ解析の“心”とは？

↓

- 目に見えること
- じっと眺めること
- 視点を変えること

どのように「視覚化」するか

# 例）アヤメの花弁形状データ



花弁の長さと幅

萼弁の長さと幅

この４変量データを考える

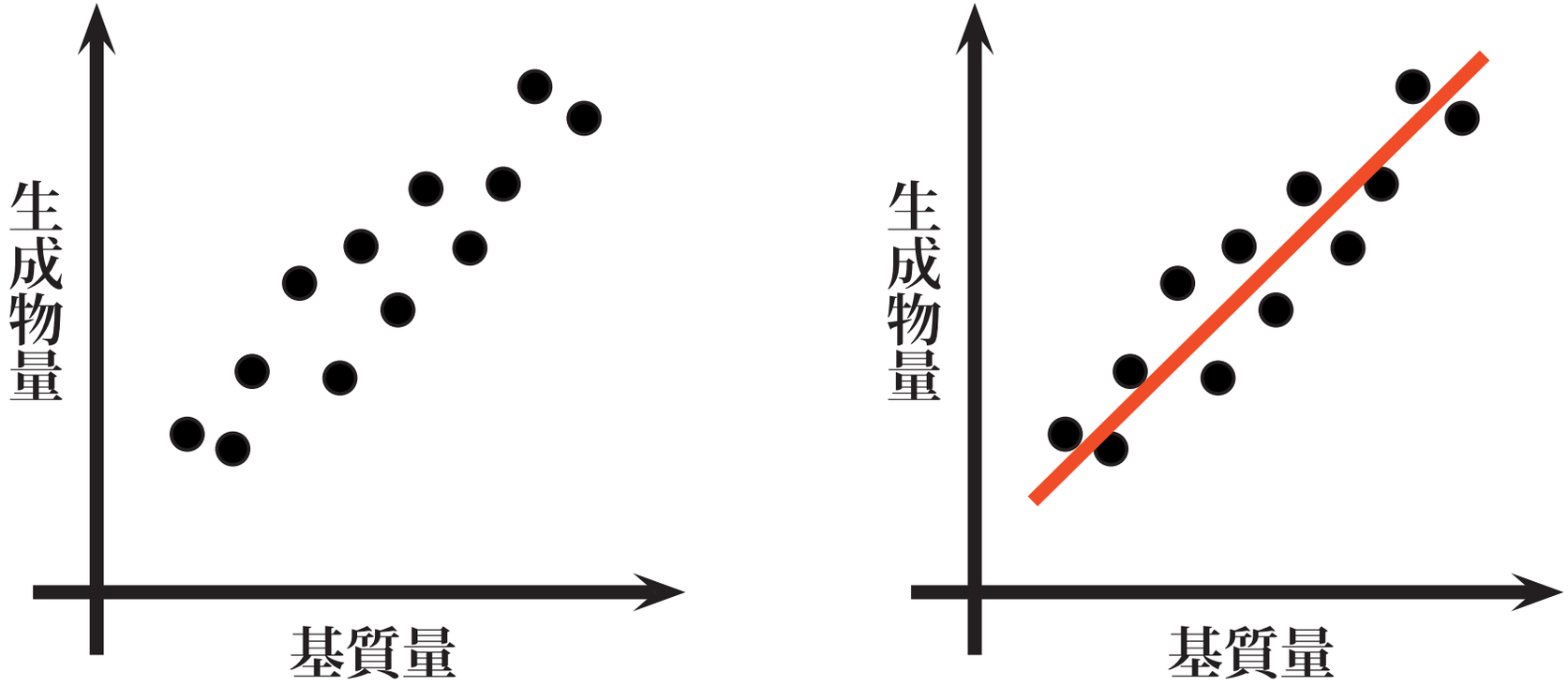
***Iris setosa***

***Iris versicolor***

***Iris virginica***

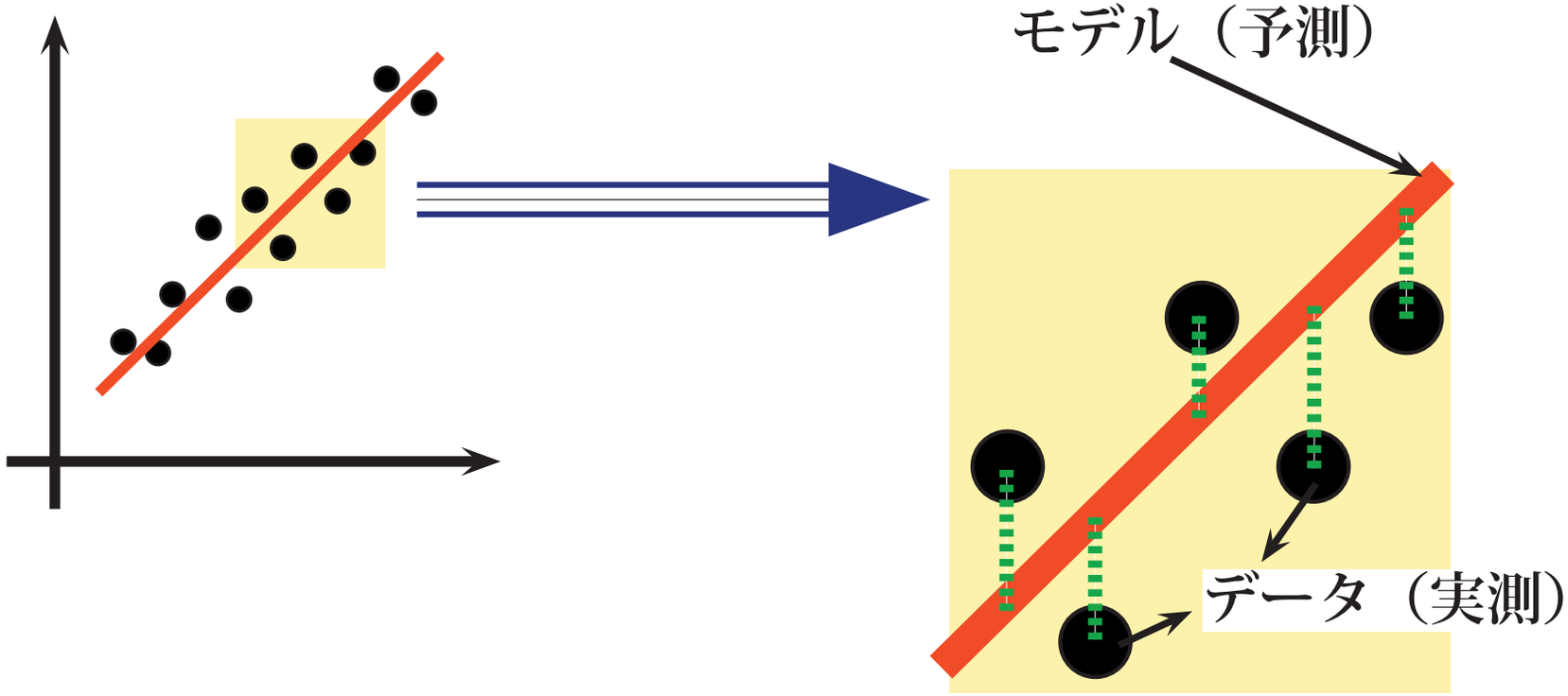
# 実験データをいきなり
# 「数式」でいじるのではなく，
# まずはじめに
# 「絵」を描いてみる．

# データをモデルで説明する

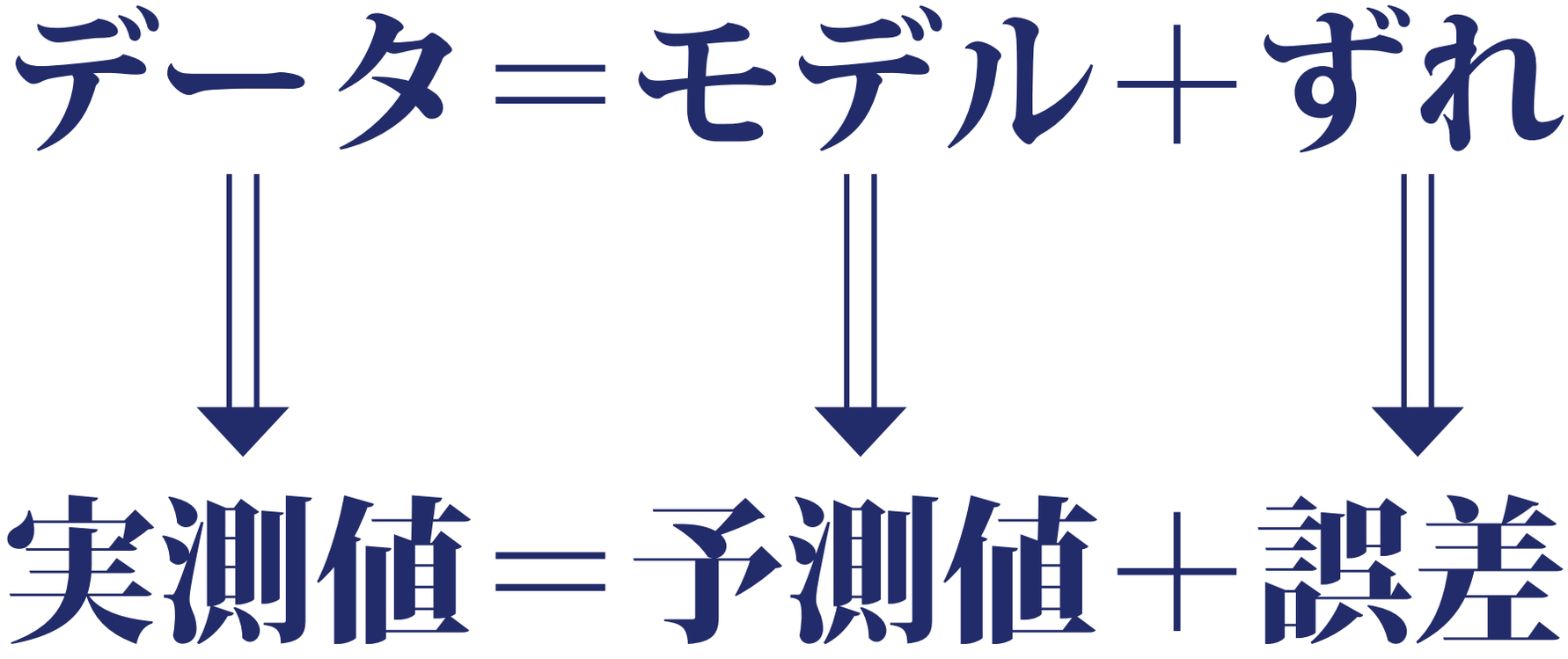



データ ⇢ モデル

# データをモデルで説明する



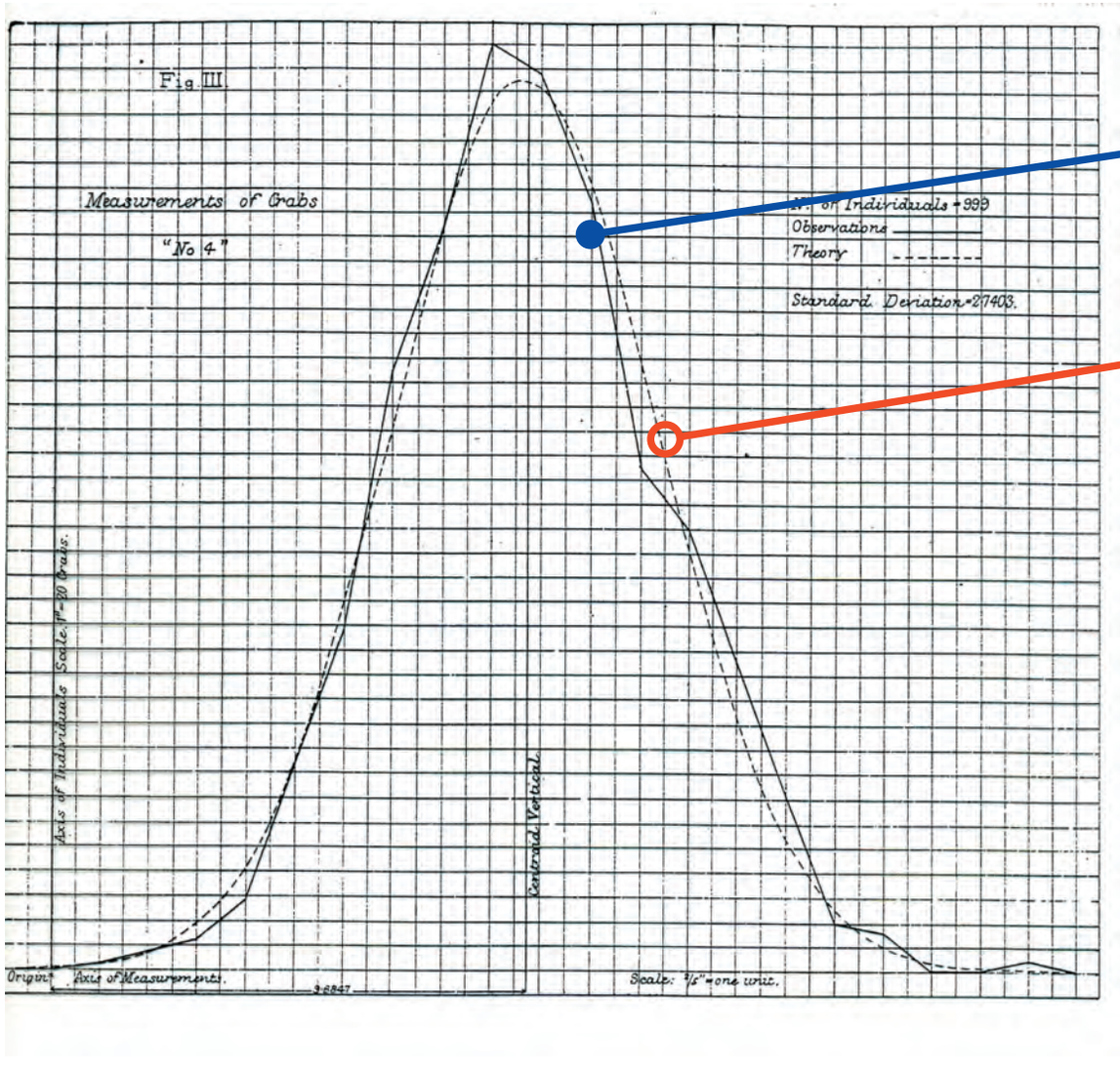

## モデルはデータから「ずれる」ものである


データ＝モデル＋ずれ
実測値＝予測値＋誤差

# パラメトリック統計学の出発点

データのもつ「誤差」の性質は,
「誤差関数」によって数式表現できる.

# パラメトリック統計学の出発点




データ（実測）

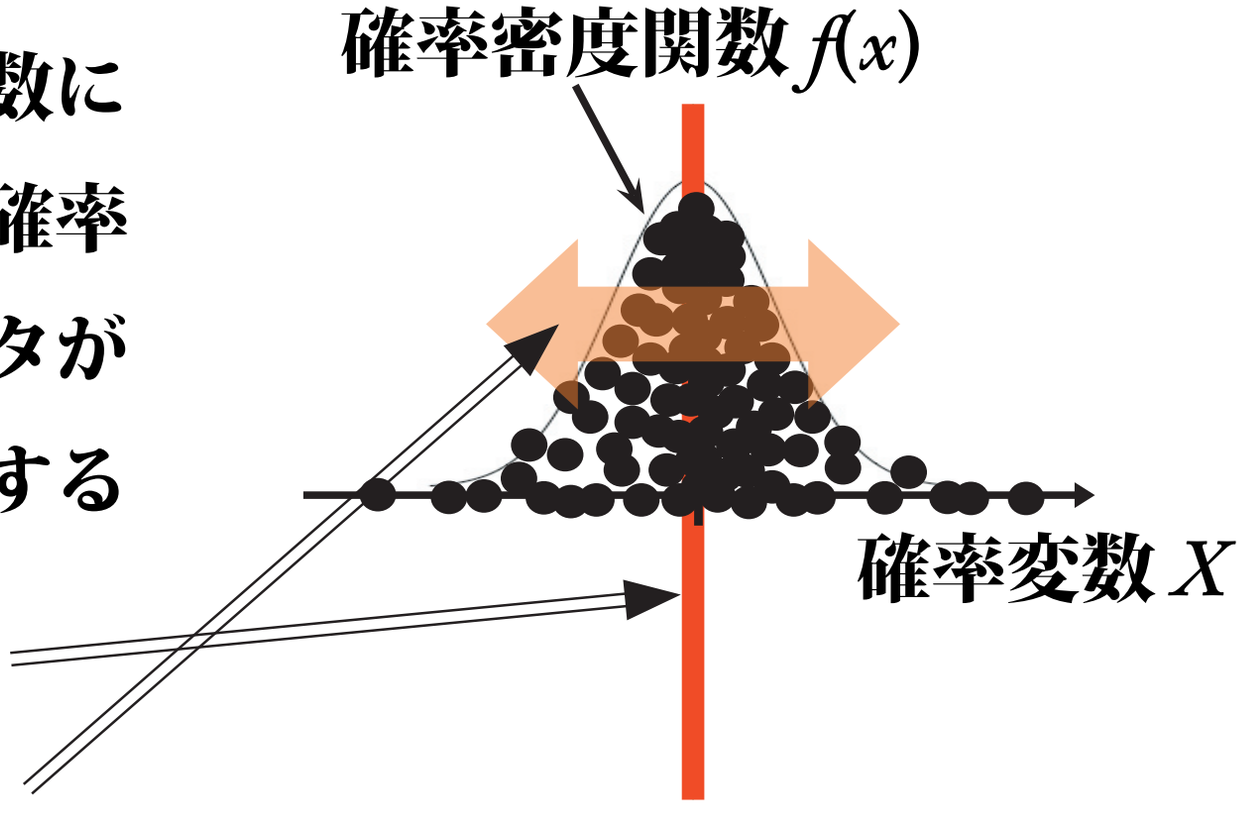
モデル（予測）

- 正規分布
- ガンマ分布
- 二項分布
- ポアソン分布

etc...

# 確率変数と確率分布

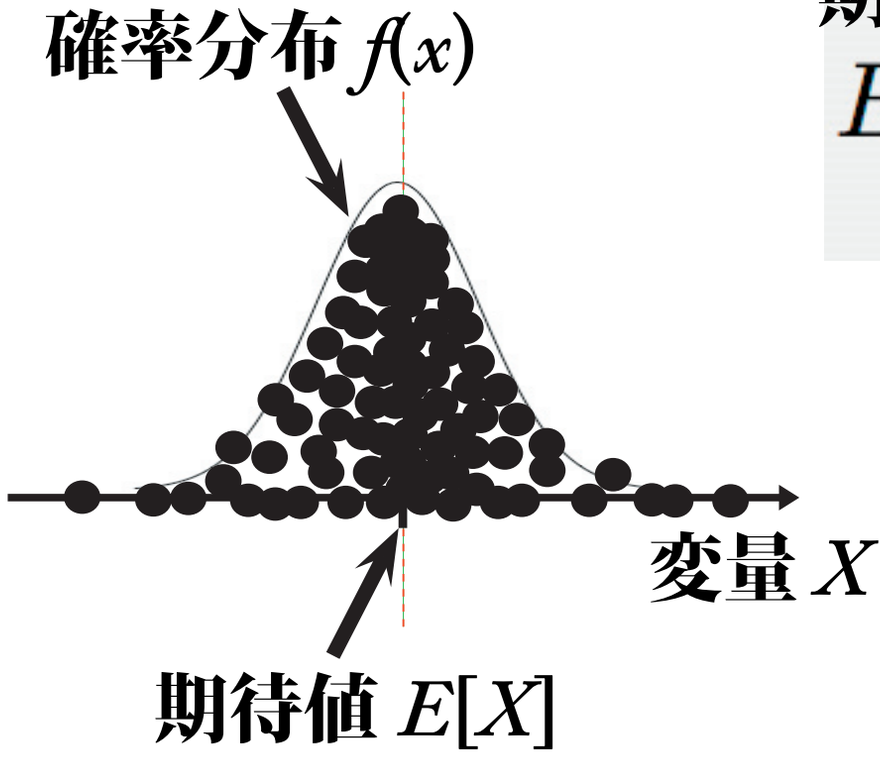
確率変数（変量）の確率分布は，確率密度関数によって定義される．確率密度関数のパラメータが確率分布の形を決定する




位置パラメータ

分散パラメータ

# 確率分布と「分散」の関係




期待値

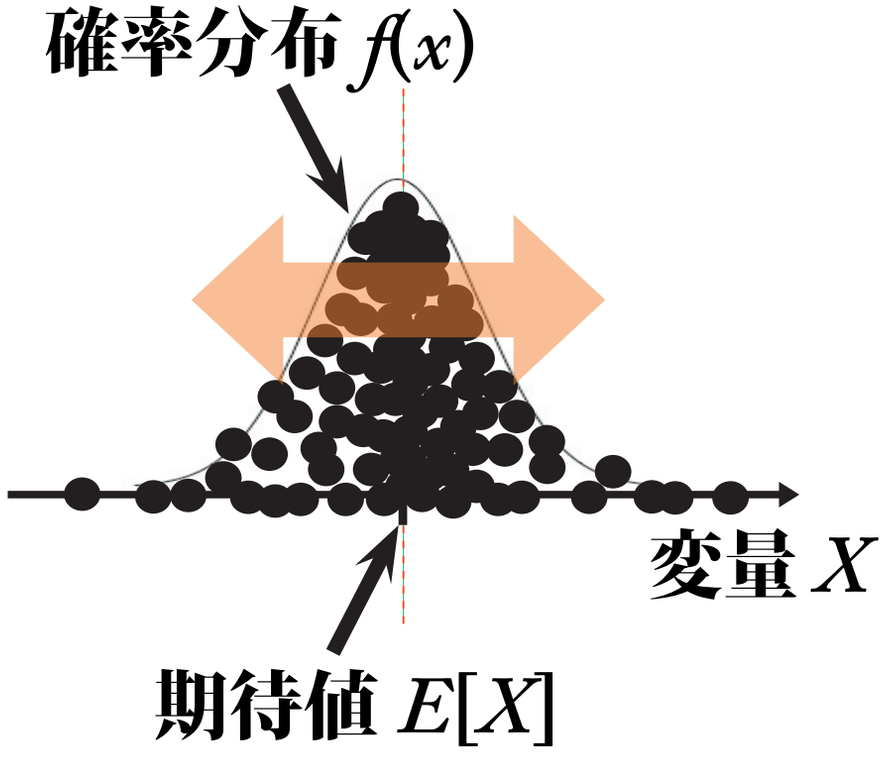
$$E[X] = \int_{-\infty}^{+\infty} x f(x)\, dx$$

変量 $X$ の値

期待値演算子 $E[.]$

# 確率分布と「分散」の関係

「偏差」とは変量の「ばらつき」の程度である.



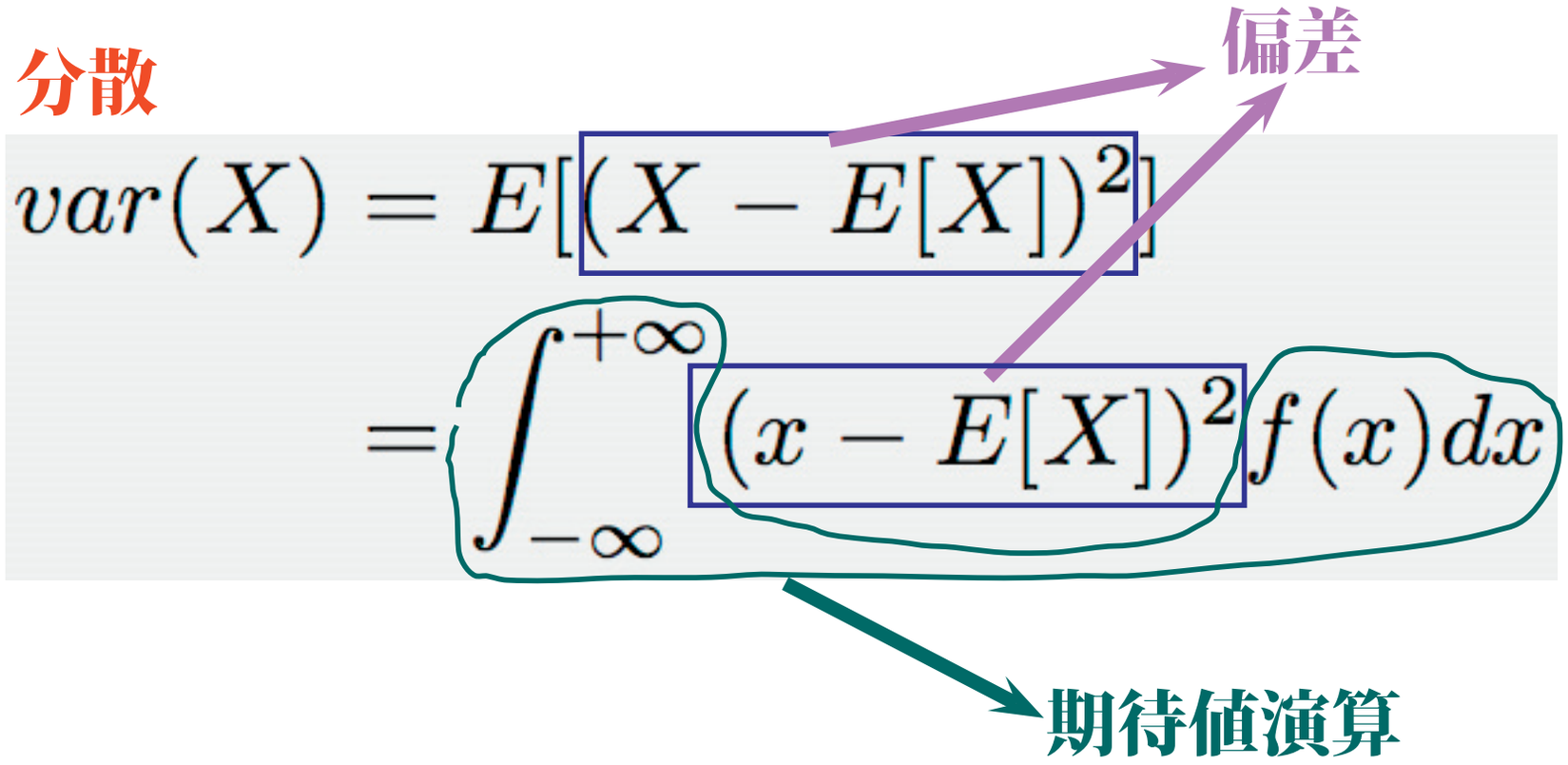

偏差 $(X - E[X])^2$

変量 $X$

$X$ の期待値
(定数)

# 確率分布と「分散」の関係

「分散」とは変量の「偏差」の期待値である.



# 確率分布と「分散」の関係

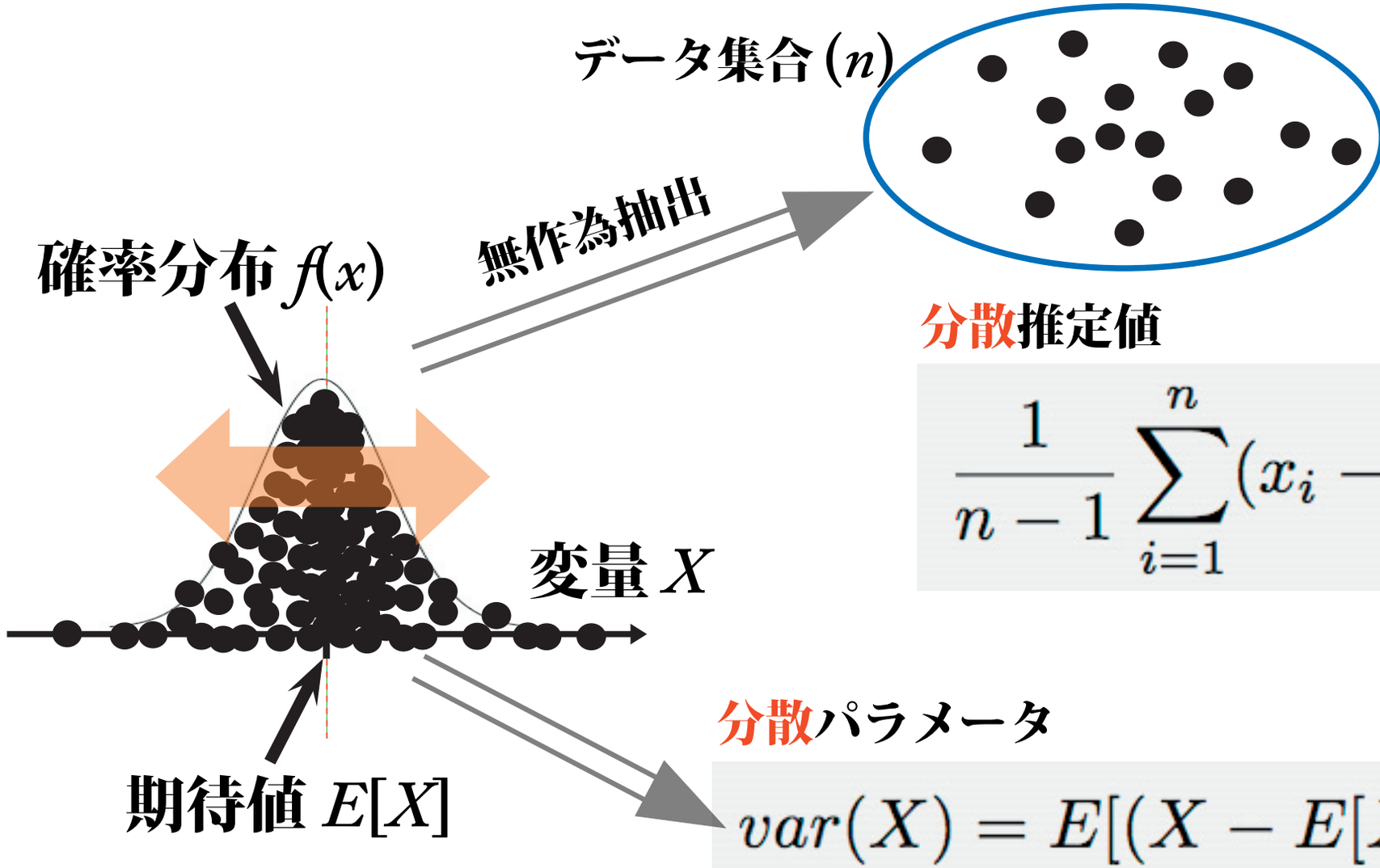
ある変量 $X$ の確率分布の密度関数が $f(x)$ で与えられ，その期待値を $E[X]$ とするとき，$X$ の「分散」$var(X)$ は：

$$var(X) = \int_{-\infty}^{+\infty} (x - E[X])^2 f(x) dx$$

ただし，

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

# 「分散」パラメータとその推定値



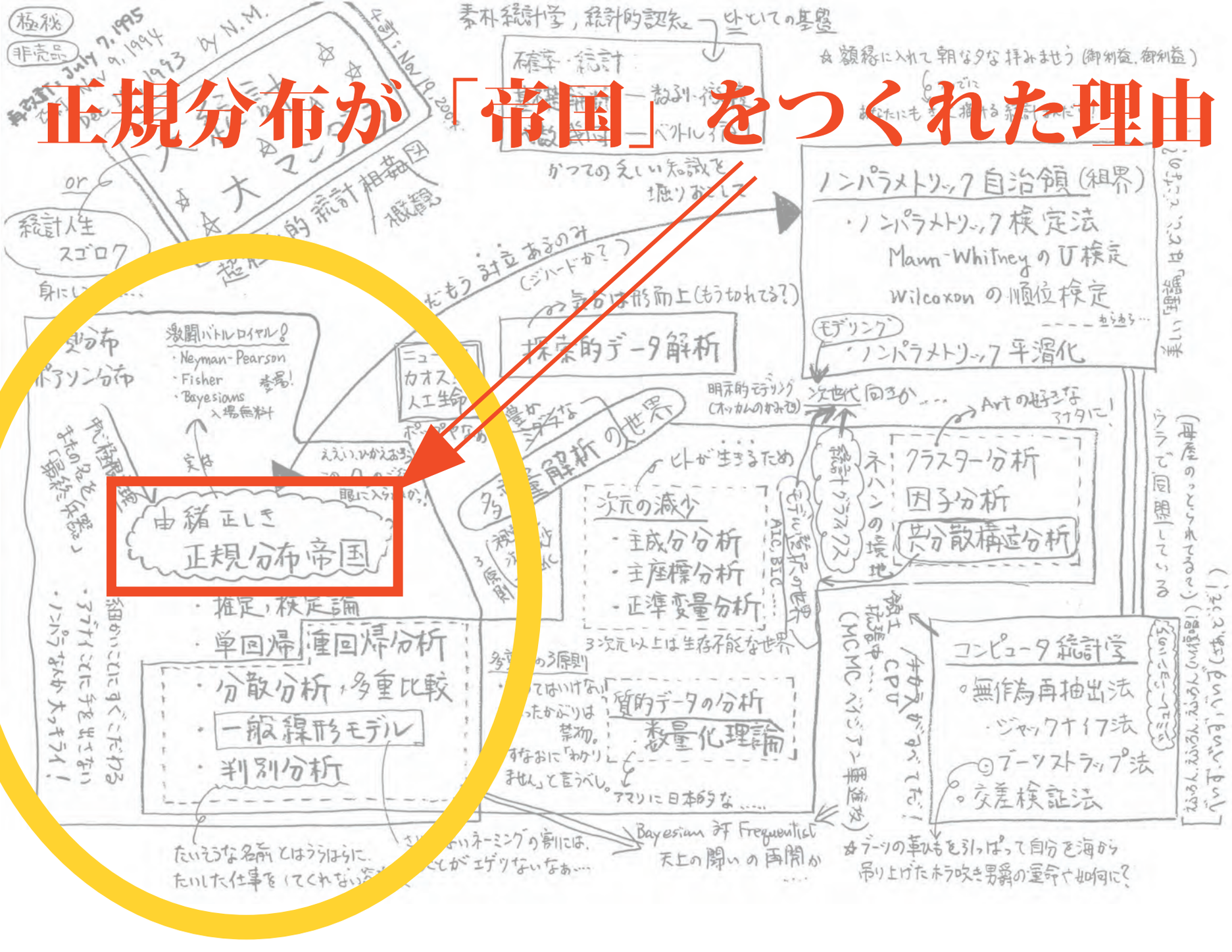

分散推定値

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

分散パラメータ

$$var(X)=E[(X-E[X])^2]$$

# 正規分布が「帝国」をつくれた理由



# 正規分布

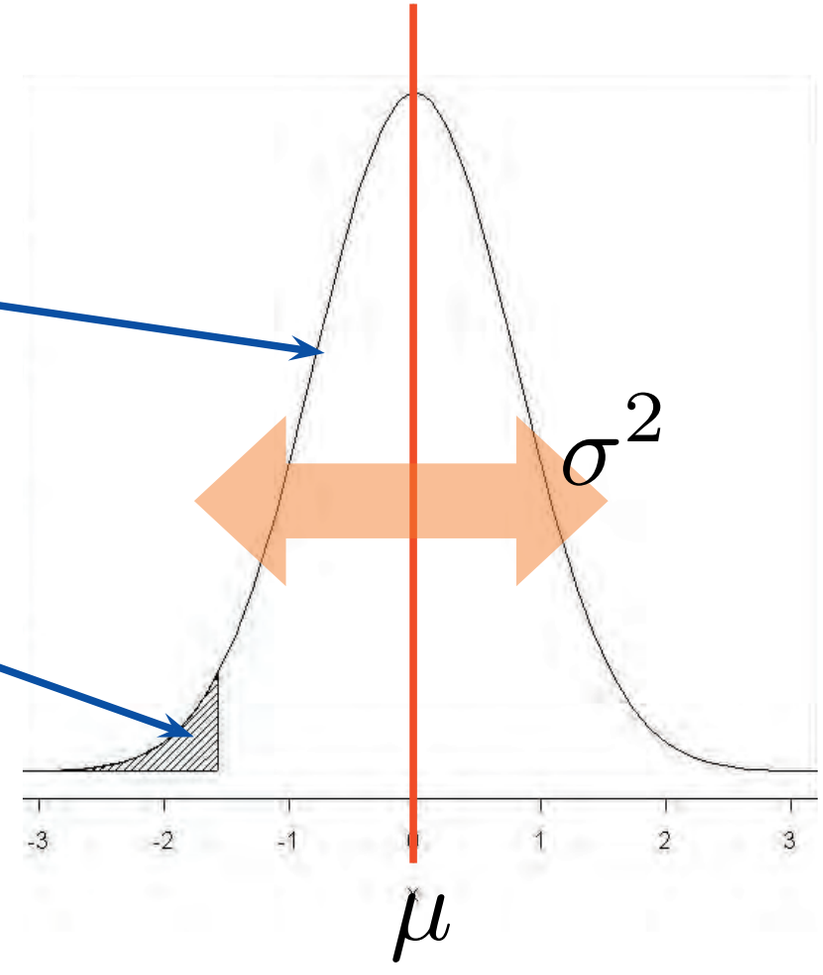
平均$\mu$，分散$\sigma^2$の正規分布 $N(\mu, \sigma^2)$の確率密度関数

## 確率密度関数

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

## 累積確率関数

$$P(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



# 正規分布

## 線形変換における正規性の保存

1) $X \sim N(\mu, \sigma^2)$ のとき, $aX + b \sim N(a\mu + b, a^2\sigma^2)$

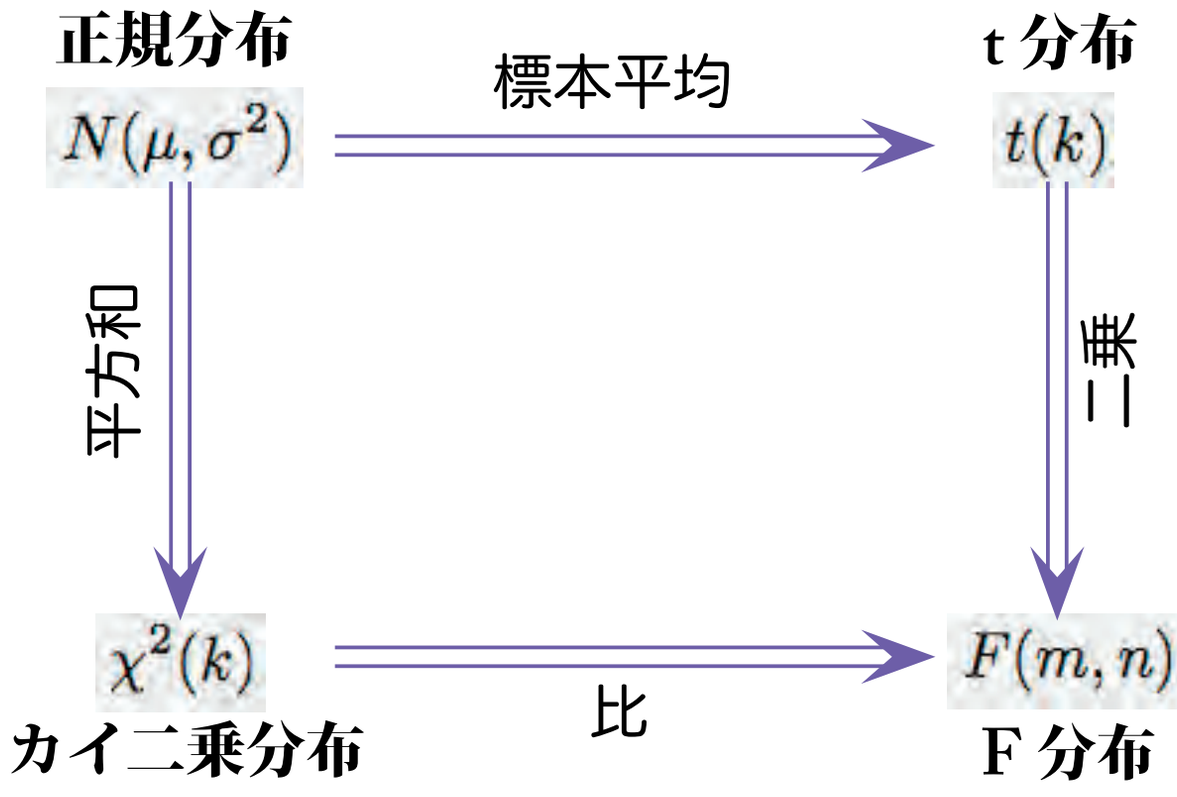
とくに, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$ を標準正規分布と呼ぶ.

2) $X_i \sim N(\mu_i, \sigma_i^2)(i = 1, 2, \ldots, n)$ のとき,

$$\sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, a_i^2 \sigma_i^2)$$

# 正規分布

## 関連する他の確率分布の導出



# 正規分布

## 関連する他の確率分布の導出

### カイ二乗分布

$$\chi^2(k) = \frac{1}{\Gamma(\frac{k}{2})}\left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$$

### t 分布

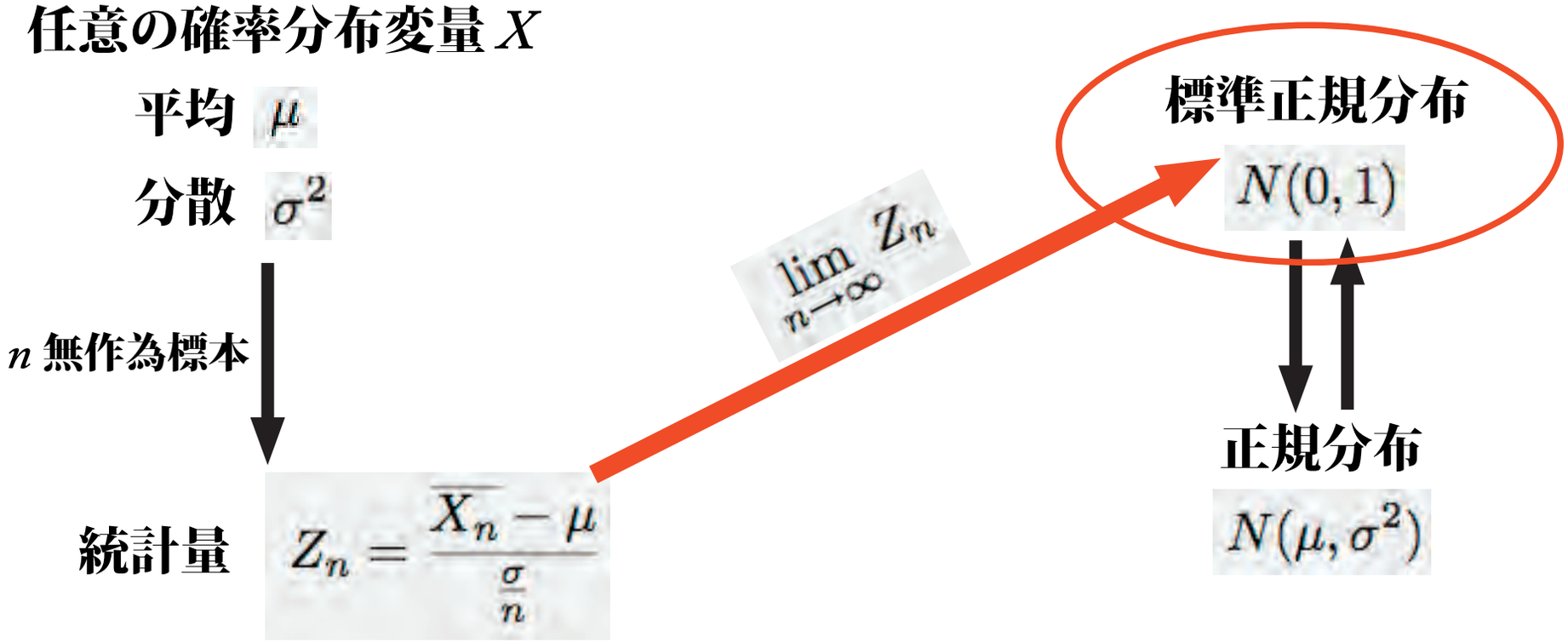
$$t(k) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}\frac{1}{\sqrt{k\pi}}\frac{1}{\left(1+\frac{x^2}{k}\right)^{\frac{k+1}{2}}}$$

### F 分布

$$F(m,n) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{\frac{m}{2}}\frac{x^{\frac{m-2}{2}}}{[1+(\frac{m}{n})x]^{\frac{m+n}{2}}}$$

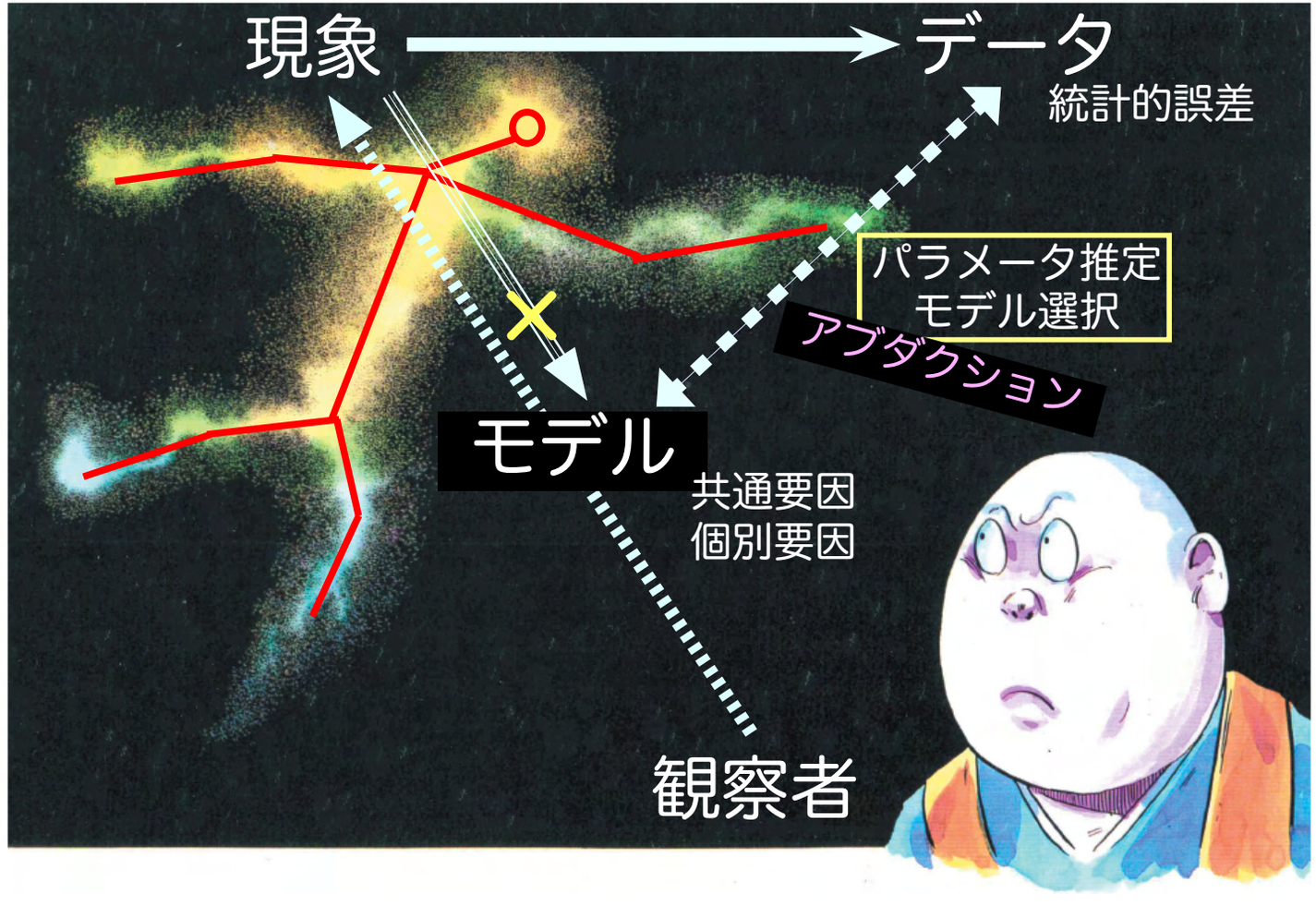
# 正規分布

## 「中心極限定理」


任意の確率分布変量 $X$
平均 $\mu$
分散 $\sigma^2$
$n$ 無作為標本
統計量 $Z_n = \dfrac{\overline{X_n} - \mu}{\frac{\sigma}{n}}$
$\lim_{n\to\infty} Z_n$
標準正規分布
$N(0,1)$
正規分布
$N(\mu, \sigma^2)$

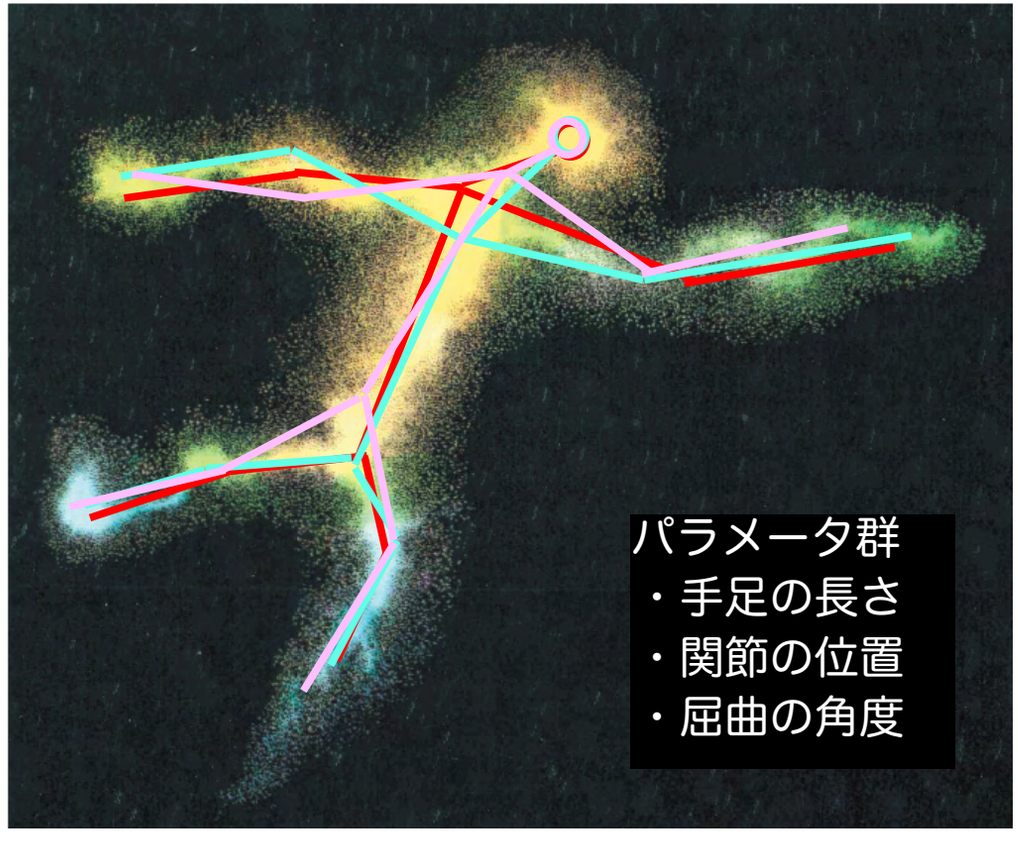
# 線形統計モデル（LM）への入り口

## モデルとデータの接点に触る



# 線形統計モデル（LM）への入り口

## モデルの構造とパラメータ



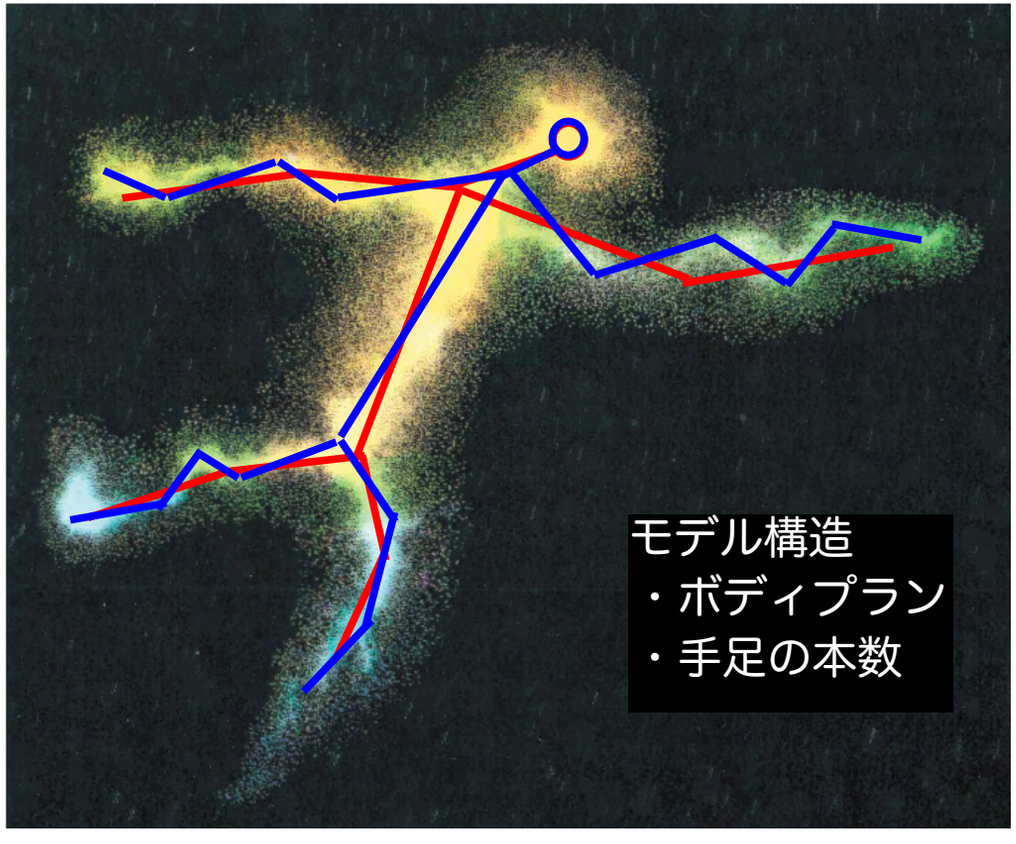

### パラメータ推定問題

モデルの「構造」を固定したとしても，パラメータ群の値は可変である．どのような基準で最適なパラメータ値を決定するのか．

計算すればなんとかなる

# 線形統計モデル（LM）への入り口

## モデルの構造とパラメータ




### モデル選択問題

モデルの基本となる「構造」そのものをどのような基準で決定すればいいのか.

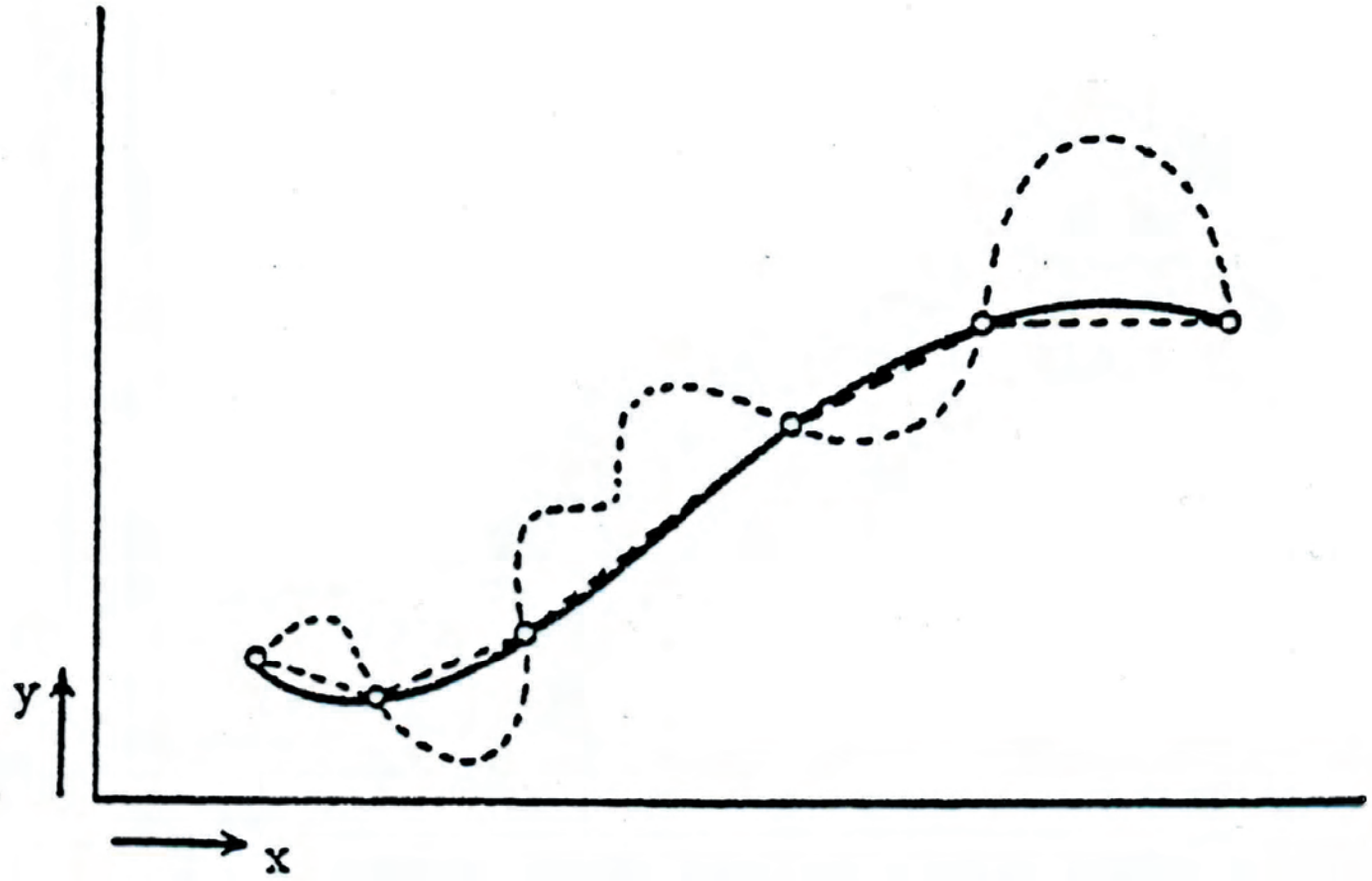
やや難しいかも・・・

# 仮説選択

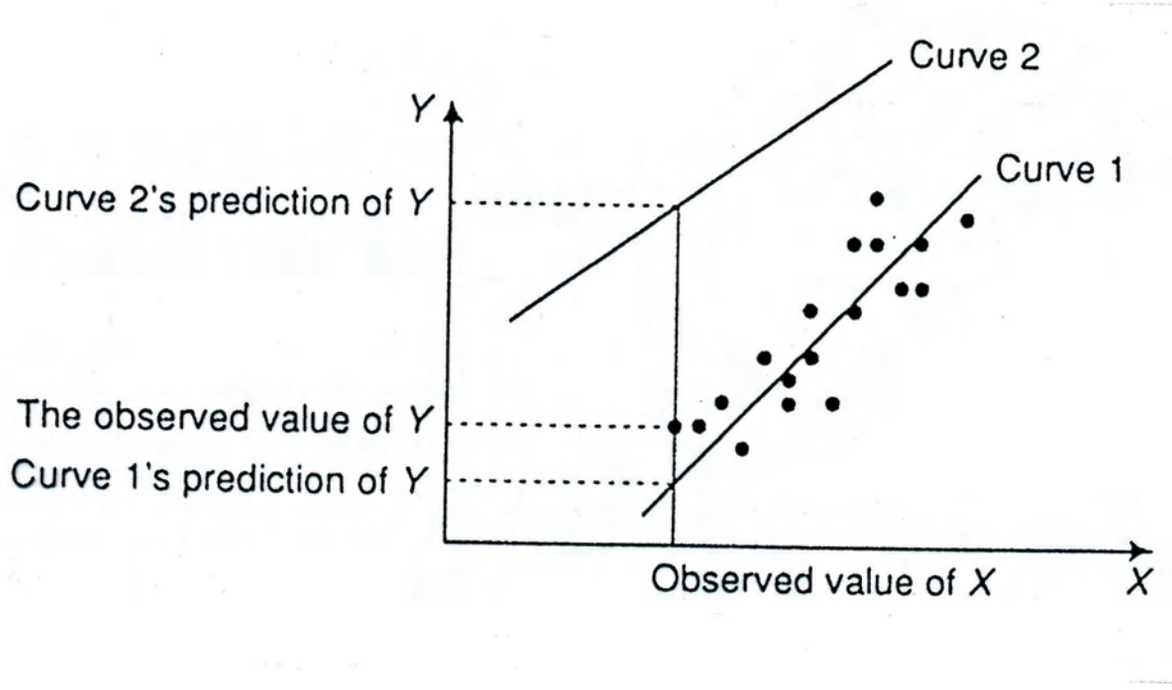
# あるいは

# モデル選択

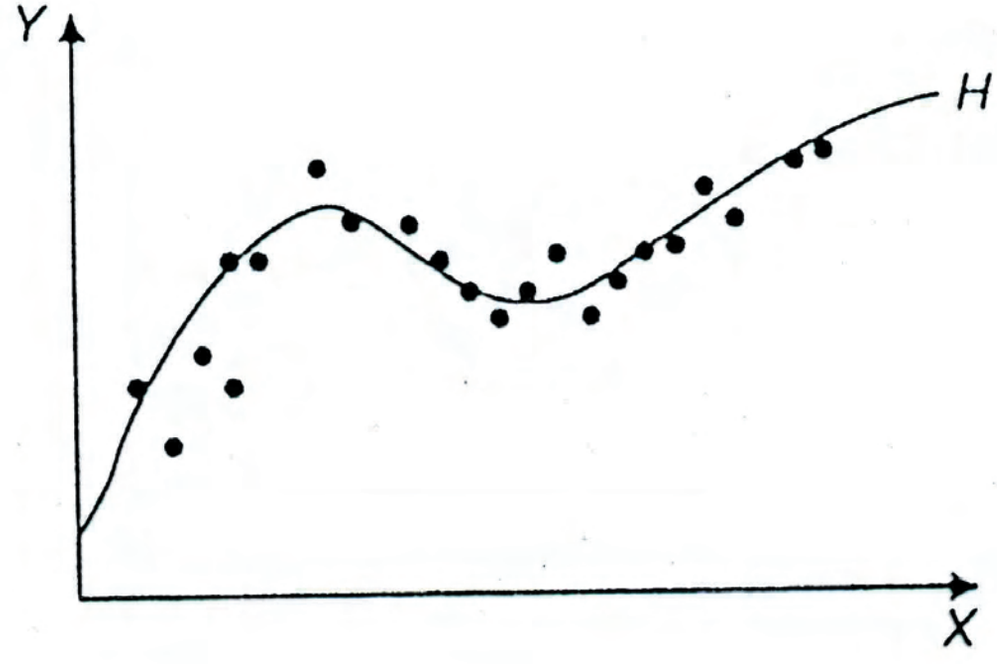
# Curve-fitting problem (1)



# Curve-fitting problem (2)



# Curve-fitting problem (3)

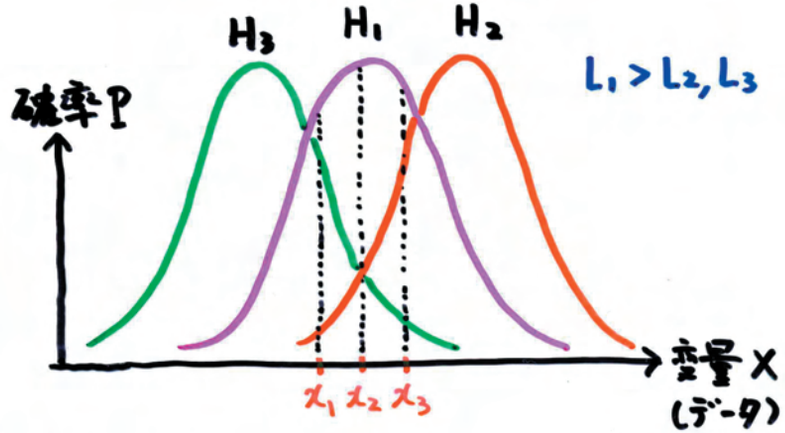


# 最尤法

# *Maximum Likelihood method*

# 尤度（Likelihood）

データが仮説に与える支持の強さ．



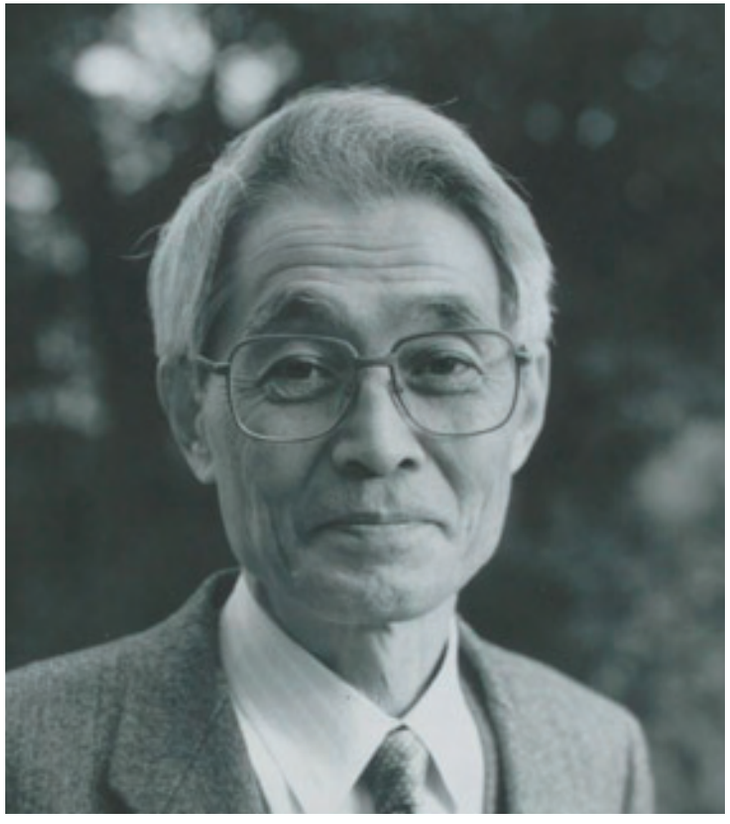

仮説 $H_1$, $H_2$, $H_3$ } 確率分布の平均値（パラメーター）が異なる．

尤度 $L = P(x_1) \cdot P(x_2) \cdot P(x_3)$
（確率積）

↓

最尤推定：尤度Lが最大となる仮説を受容する．

複雑なモデルは「最良の説明」を必ずしも与えないだろう．単純なモデルの方がこの点で勝っている可能性がある．



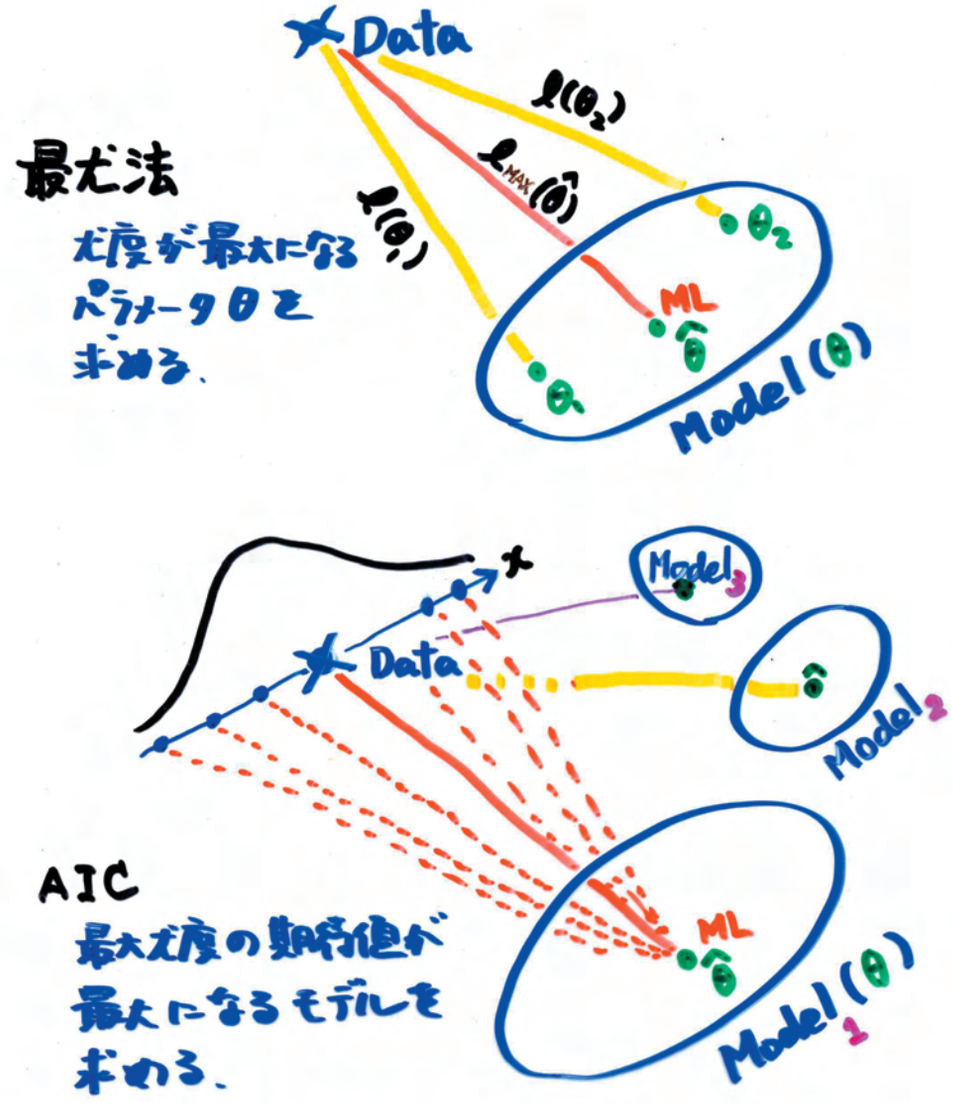
良いモデルは
「予測確度」
(predictive accuracy)
が大きい

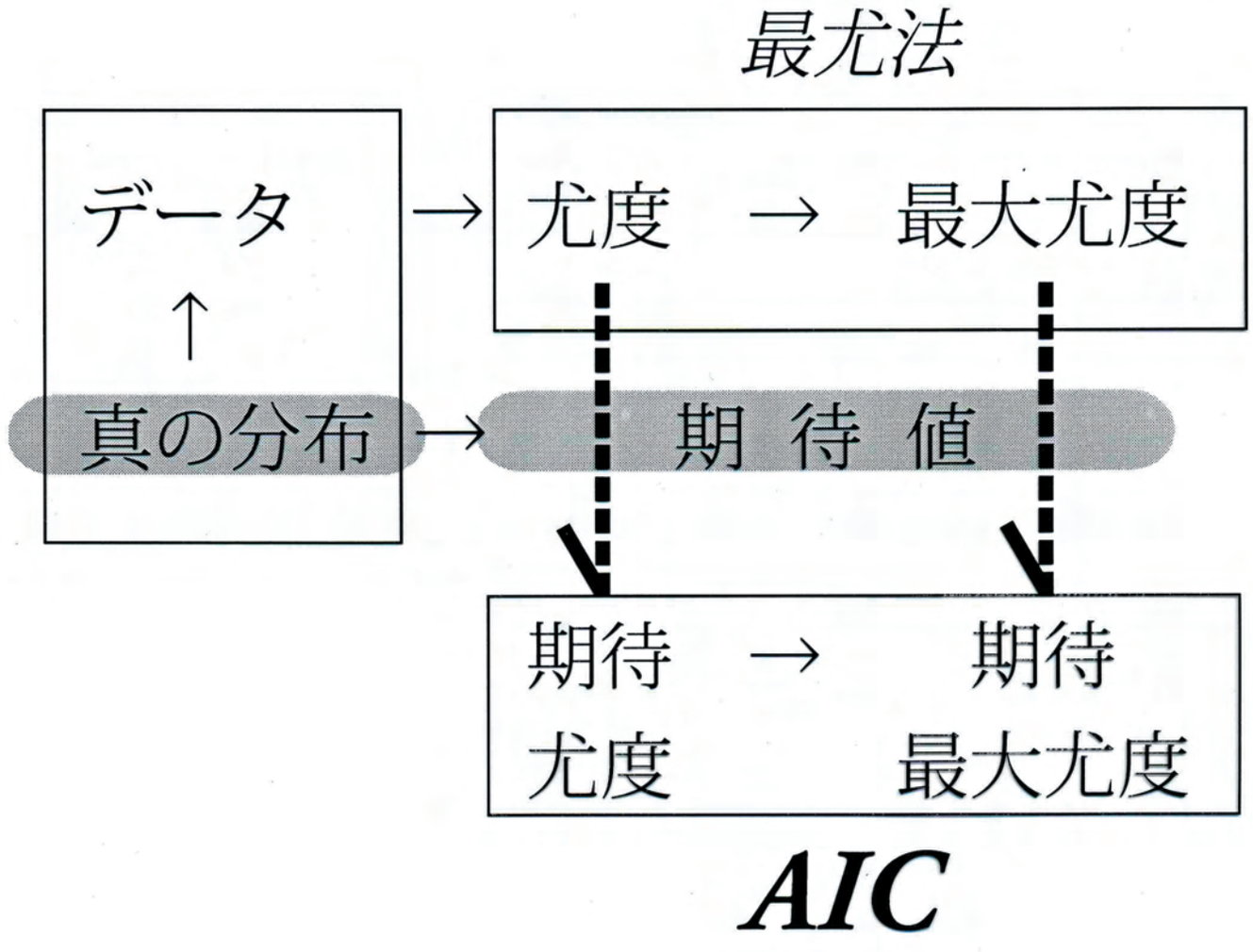
［赤池弘次］

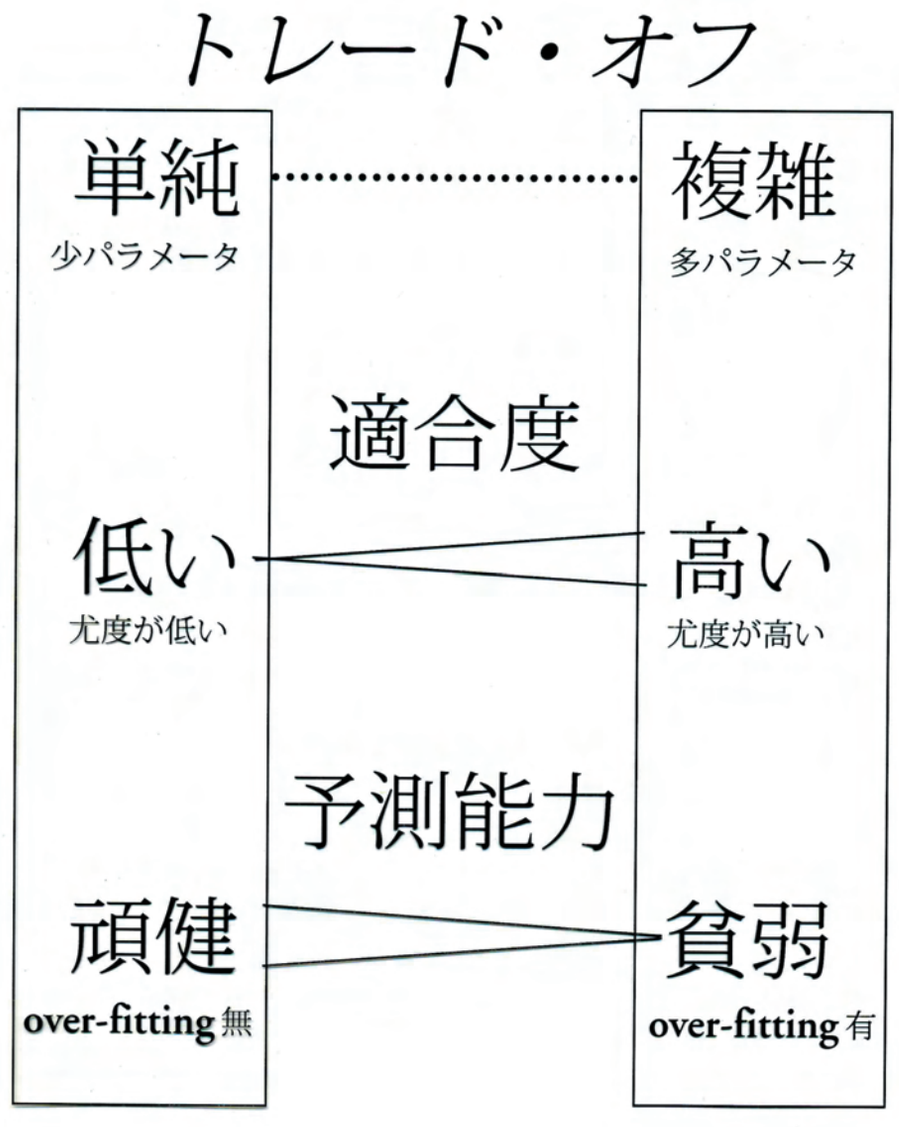
第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法

# *AIC*

## （赤池情報量基準）


最尤法
尤度が最大になる
パラメータθを
求める。
Data
ML
Model(θ)
AIC
最大尤度の期待値が
最大になるモデルを
求める。
Data
ML
Model(θ)


最尤法
データ
↑
真の分布
尤度 → 最大尤度
期 待 値
期待 → 期待
尤度 最大尤度
AIC


トレード・オフ
単純
少パラメータ
複雑
多パラメータ
適合度
低い
尤度が低い
高い
尤度が高い
予測能力
頑健
over-fitting無
貧弱
over-fitting有

# モデルの予測確度

＝

# 期待尤度－パラメータ数

（適合度）　　（複雑度）

↓

# *AIC*

（赤池情報量基準）

# 線形モデルを鍛えあげる

## 線形モデル（LM）

データ　総平均

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad (\epsilon_{ij} \sim N(0, \sigma^2))$$

処理効果　誤差項　誤差は正規分布をする

通常の「線形モデル」の仮定

- 処理効果はある実数（「パラメータ値」）である.
- 誤差項は独立かつ同一の正規分布にしたがう.
  （「正規性」と「等分散性」）

窮屈でしかたがない！

# 線形モデルを鍛えあげる

## 線形モデル（LM）

データ　総平均

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad (\epsilon_{ij} \sim N(0, \sigma^2))$$

処理効果　誤差項　誤差は正規分布をする

通常の「線形モデル」の仮定

- 処理効果はある実数（「パラメータ値」）である.
- 誤差項は独立に正規分布にしたがう.
  （「正規性」と「等分散性」）

これらの仮定を緩めよう（一般化）

# 線形モデルを鍛えあげる

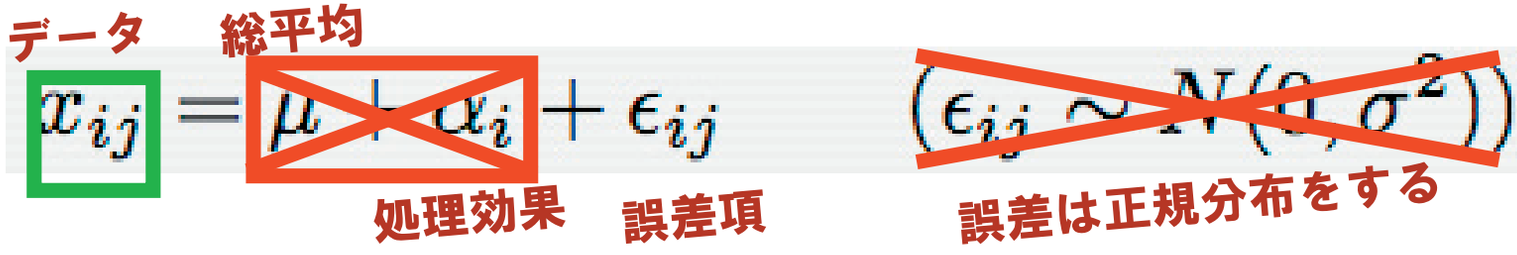
## 一般化線形モデル（GLM）

何が一般化されたのか？

- 誤差項が正規分布でなくてもよい.
- データのある関数変換（「リンク関数」）が線形モデルで記述できればよい

# 線形モデルを鍛えあげる

## 一般化加法モデル（GAM）




さらなる一般化とは？

- 予測関数は単一の式で表示される関数である必要すらなく，スプライン関数によって区分的に構築された任意に関数（ただし加法性はある）でもかまわない．

# 線形モデルを鍛えあげる

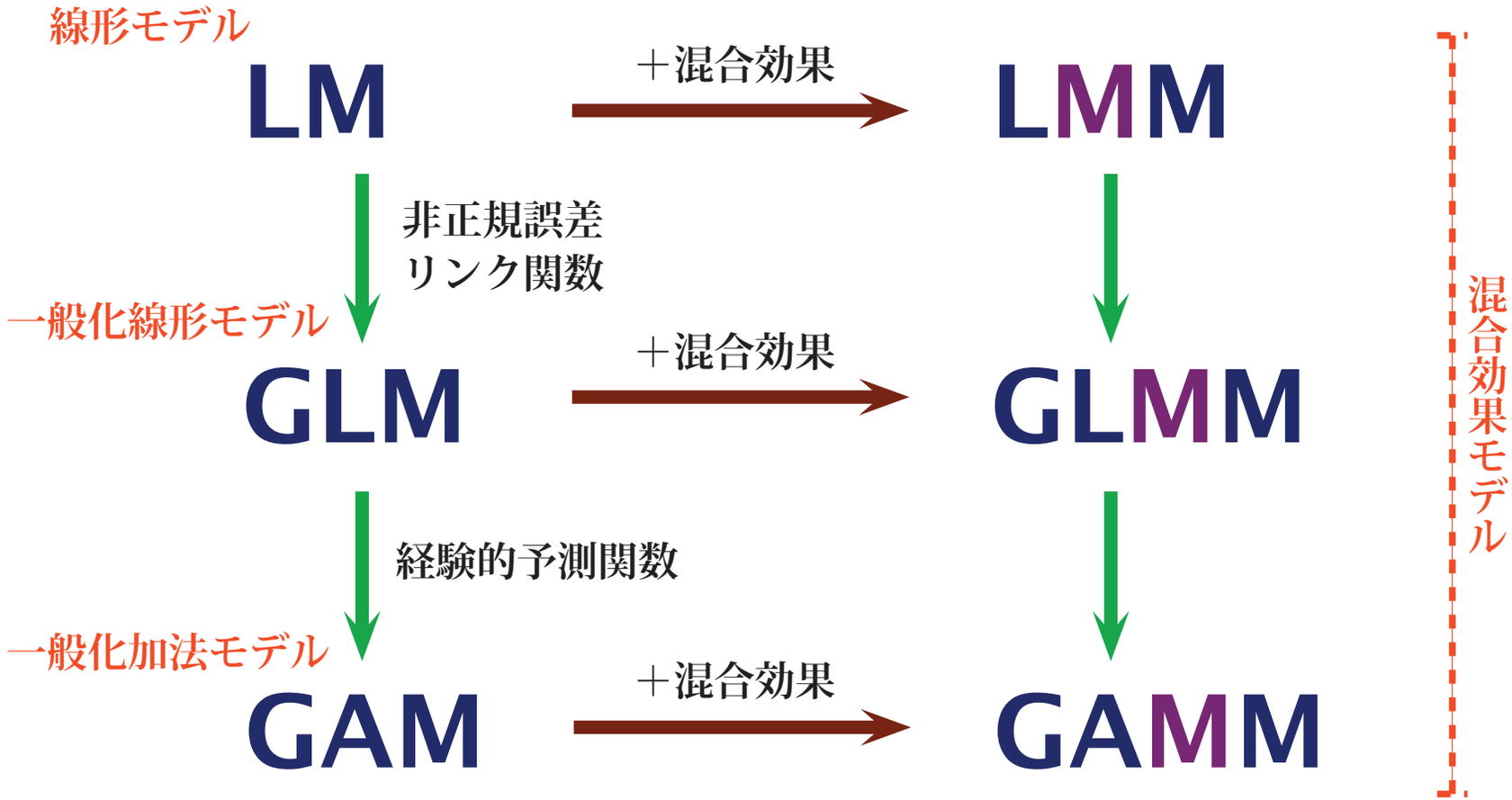
## 混合効果モデル（mixed-effect model）

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad (\epsilon_{ij} \sim N(0, \sigma^2))$$

処理効果

要因の「処理効果」は，ある実数の「パラメータ値」という定数として「固定効果」（fixed effect）をもつと仮定してきた．しかし，その処理効果が別の確率分布にしたがう“変量”とみなして，「ランダム効果」（random effect）をもつと仮定すれば，より細かい変動や個体差をもモデリングできるだろう．

# 線形モデルを鍛えあげる



# 母集団からのサンプリング

標本から計算された「統計量」はどれくらいのバラツキを伴っているのか？

リンゴ 10 個（標本）

統計量（サイズ，糖度，. . . ）

サンプリング

推定・検定

リンゴ園（母集団）

統計量に伴う「誤差」を標本データから推定すればよい

# 母集団からのサンプリング

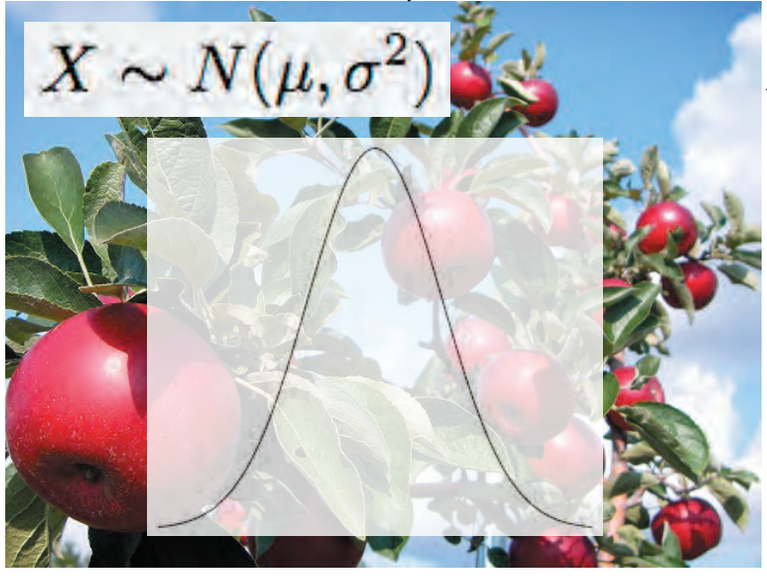
母集団が正規分布するならば，標本平均もまた正規分布をする

標本

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

サンプリング

推定・検定




母集団

$$var(\overline{X}) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n(n-1)}$$

標本平均の分散推定値は標本データそれ自体から計算できる

# 母集団からのサンプリング

## なぜ苦しまねばならないのか？

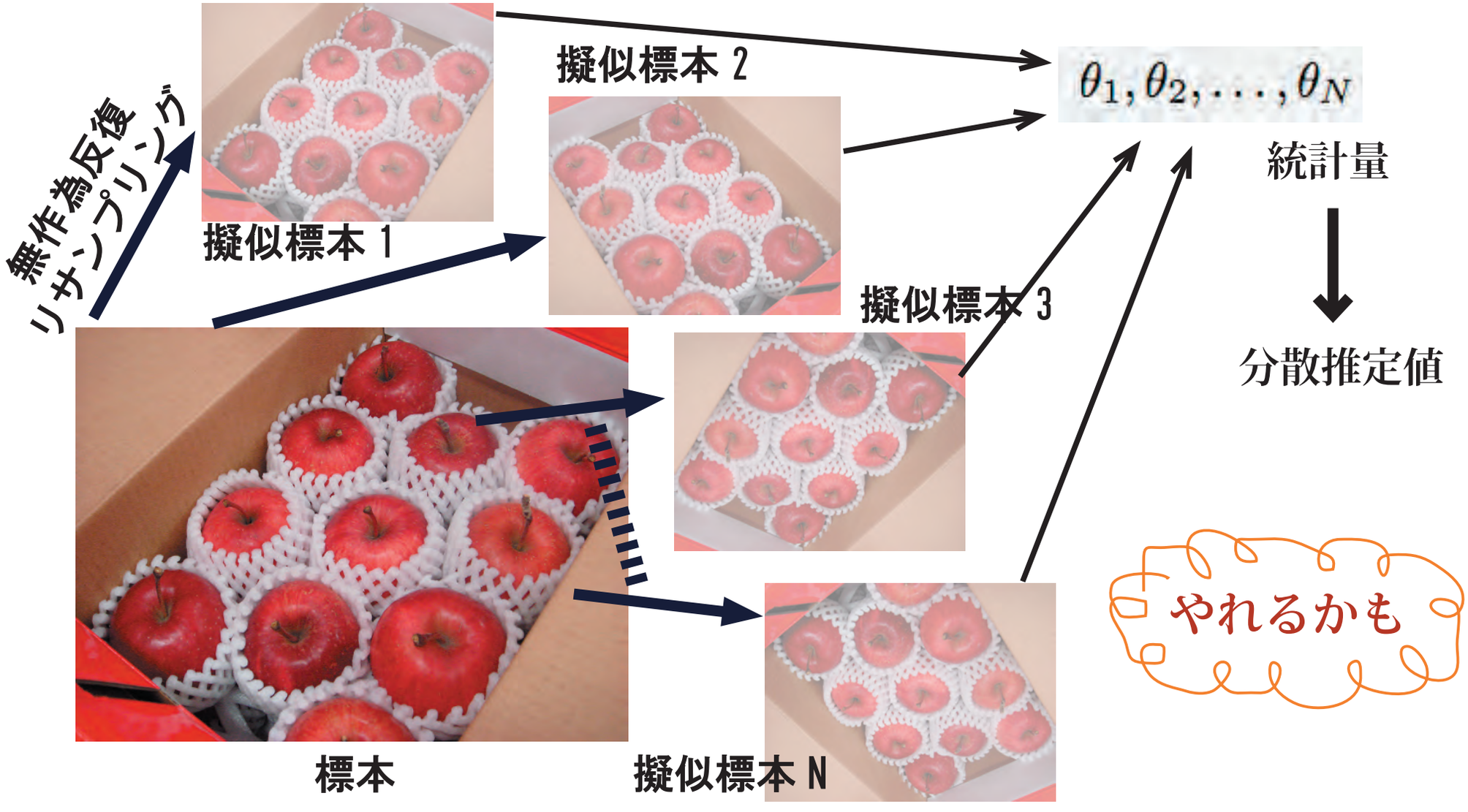
統計量の誤差の推定値は，正規分布に従う母集団のもとでの単純な統計量の場合を除いては，一般に解析的に求めることはできない．したがって，そのような制約を緩める数値的方法が必要になる．

**リサンプリング統計手法（resampling methods）**

1) ブーツストラップ（bootstrap）
2) ジャックナイフ（jackknihe）
3) モンテカルロ（Monte Carlo）

# データからの“らくらく”リサンプリング

## 標本を仮想的母集団とみなしてしまおう



# ブーツストラップ

## 重複を許して無作為同数リサンプリングを反復する

データ = 仮想母集団

$=$

$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ (元)

1

$X_1, X_1, X_3, X_4, X_6, X_7, X_7, X_8, X_9, X_{10}$

2

$X_2, X_3, X_5, X_5, X_6, X_7, X_8, X_8, X_{10}, X_{10}$

N

$X_1, X_3, X_3, X_4, X_5, X_7, X_7, X_8, X_9, X_{10}$

同数の標本抽出

$\theta_1, \theta_2, \ldots, \theta_N$

各反復から計算された統計量の集合

$$var(\hat{\theta}_B) = \frac{\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}{(N-1)}$$

分散のブーツストラップ推定値

# ジャックナイフ

## 重複を許さず無作為削除リサンプリングを反復する

データ = 仮想母集団

$=$

$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ (元)

1

2

N

一つ削って標本抽出

$X_1, \overset{v}{X_3}, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$

$X_1, X_2, X_3, \overset{v}{X_4}, X_6, X_7, X_8, X_9, X_{10}$

$\overset{v}{X_2}, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$

$\theta_1, \theta_2, \ldots, \theta_N$

各反復から計算された統計量の集合

$$var(\hat{\theta}_J) = \frac{\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}{(N-1)}$$

分散のジャックナイフ推定値

# パラメトリック・ブーツストラップ
# ~~モンテ・カルロ~~

## パラメータ推定値に基づくデータ生成シミュレーション

データ = 仮想母集団

$=$ $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ (元)

推定値 $\hat{\mu}, \hat{\sigma}^2$

$N(\hat{\mu}, \hat{\sigma}^2)$

モデル

擬似データ生成

$\theta_1, \theta_2, \ldots, \theta_N$

各反復から計算された統計量の集合

$$var(\hat{\theta}_B) = \frac{\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}{N(N-1)}$$

分散のシミュレーション推定値

母集団が従う統計モデル（確率分布や線形モデル）を仮定し，そのパラメータを元データから推定する．そのパラメータ推定値をもつモデルから擬似データの生成シミュレーションを反復試行し，その結果から統計量の分散を推定する．