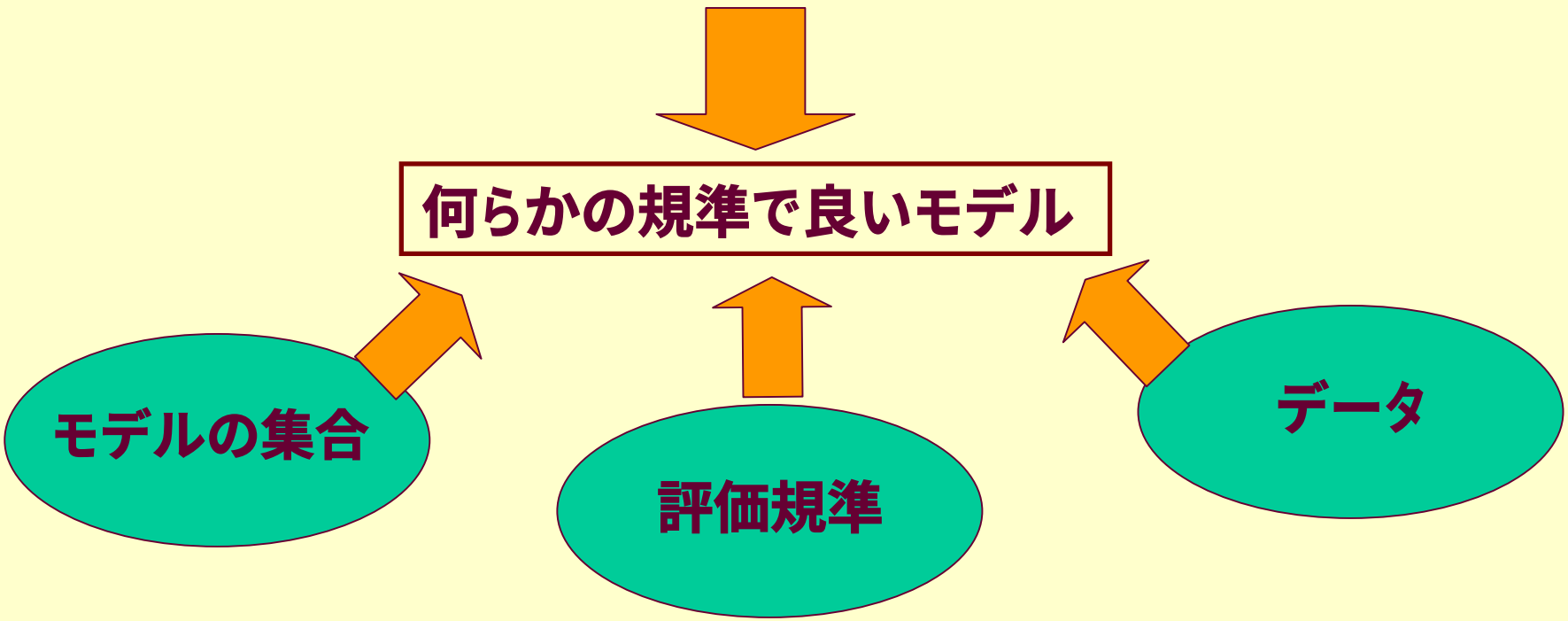
# 統計的モデル選択

## - データが選ぶ良いモデルとは？ -

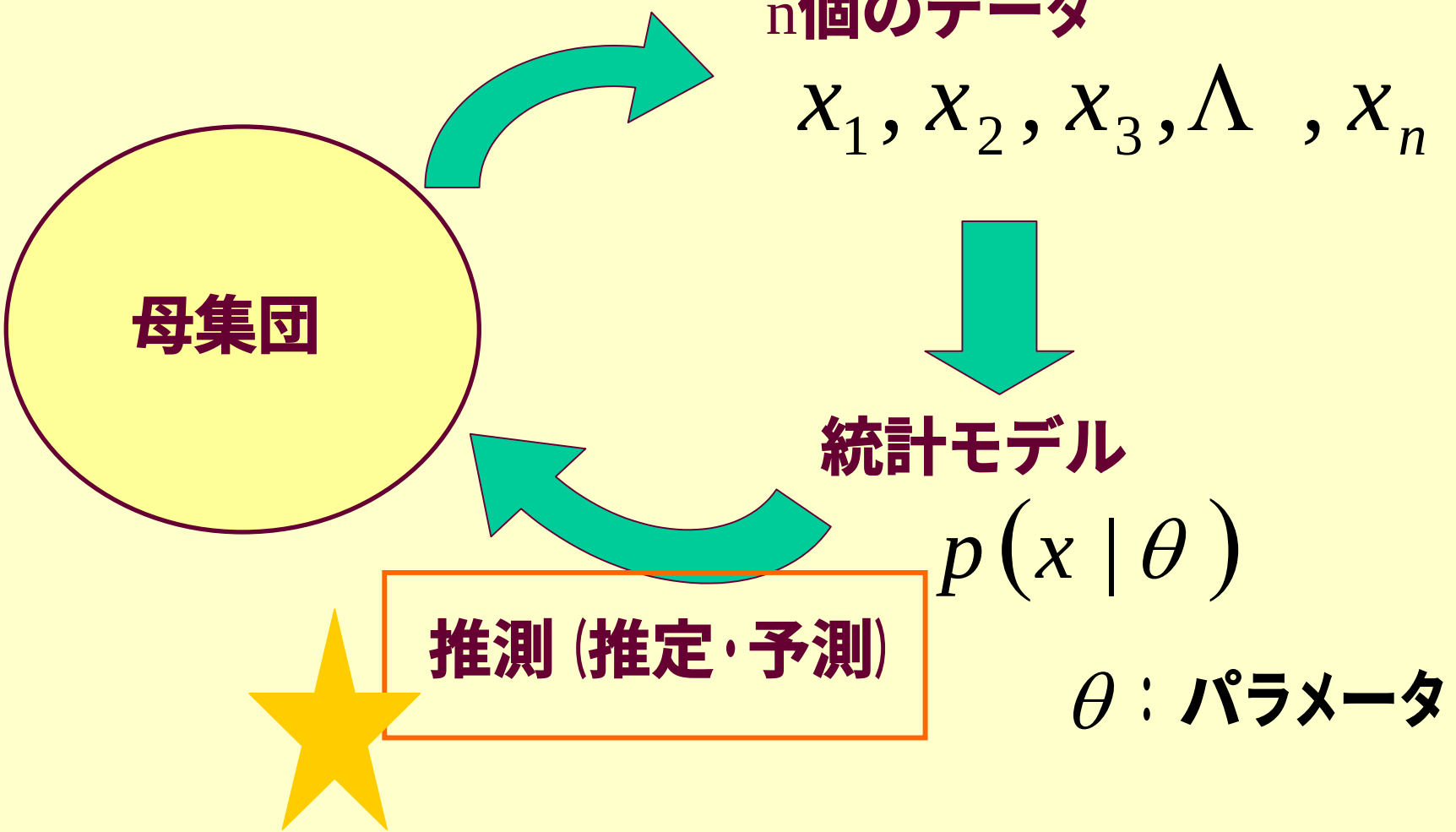
東京大学 大学院 工学系研究科
後藤 正幸

# 統計的モデル選択とは

得られたn個のデータから、良い
統計モデルを選ぶこと



# 統計的推測の原理


n個のデータ
$x_1, x_2, x_3, \Lambda\ , x_n$
母集団
統計モデル
$p(x \mid \theta)$
推測（推定・予測）
$\theta$：パラメータ

# 推定と予測

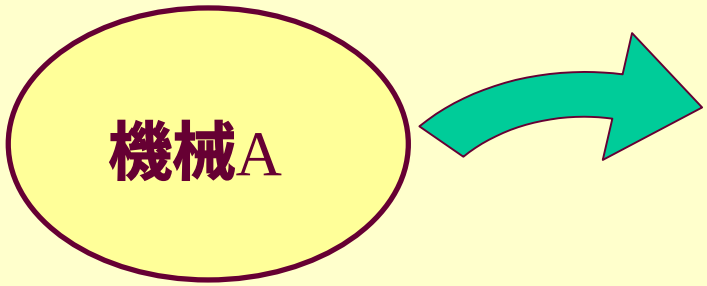
**統計的推定：**

データから、母集団の確率構造に関する何らかの値を推定すること

**統計的予測：**

過去のデータと統計モデルを使って、母集団から次に出てくるデータを予測すること

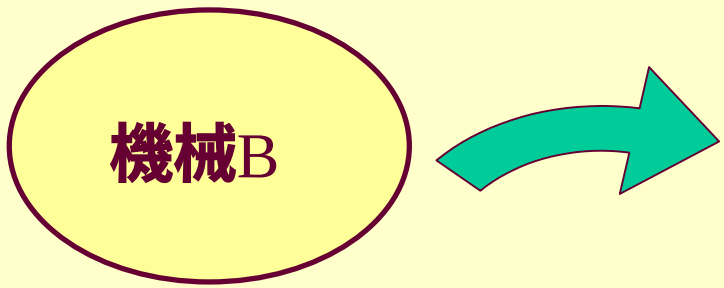
# 簡単な例題




100, 105, 95, 98, 102

算術平均 :100

平均値：$\mu_1$




106, 111, 101, 104, 108

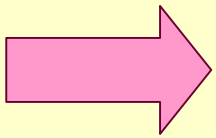
算術平均 :106

平均値：$\mu_2$

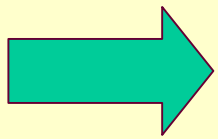
# 2つの仮説

**仮説1：**

**2台の機械で平均は同じ：** $\mu = \mu_1 = \mu_2$



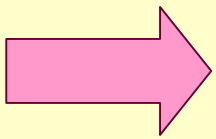
推定量 ：$\hat{\mu} = 103.0$



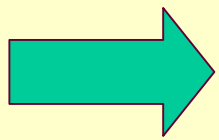
**統計モデル1**

**仮説2：**

**2台の機械で平均は異なる：** $\mu_1 \neq \mu_2$

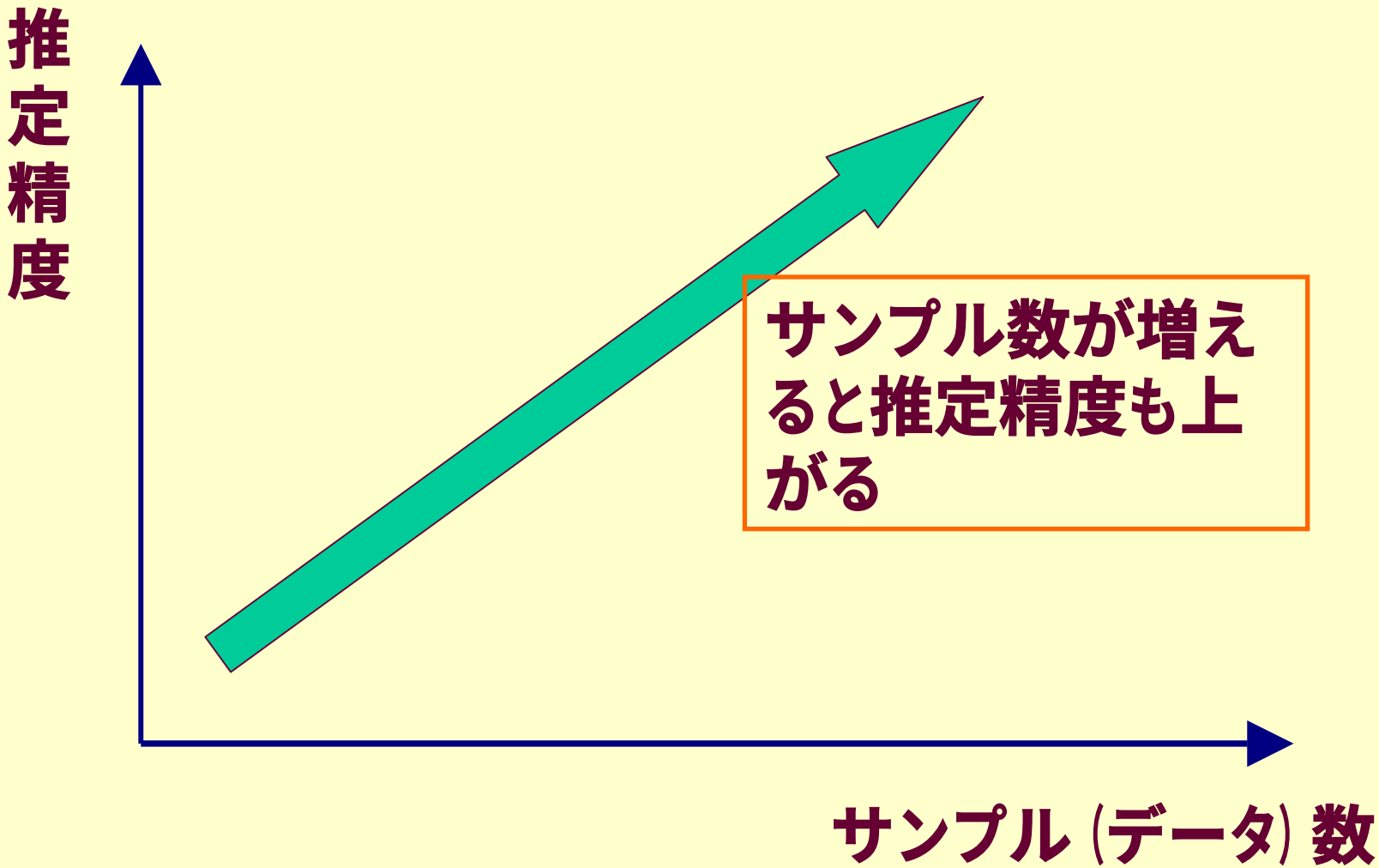


推定量：$\hat{\mu}_1 = 100.0, \quad \hat{\mu}_2 = 106.0$



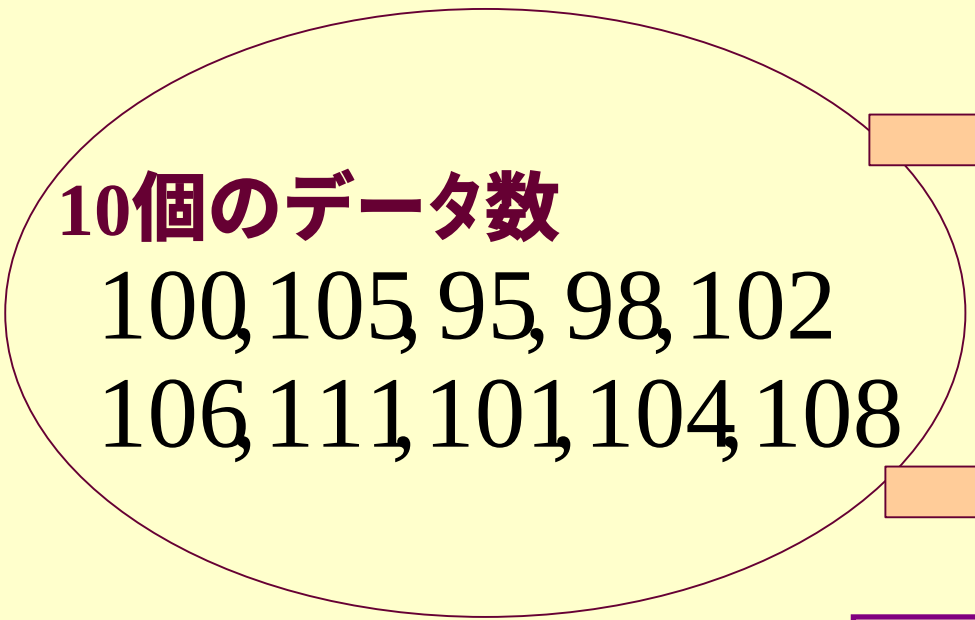
**統計モデル2**

# サンプル数と推定精度



# 2つのモデルと推定精度

- 同じサンプル数のとき、モデル1と2では？




モデル1

$\hat{\mu} = 103.0$

データ10個から推定

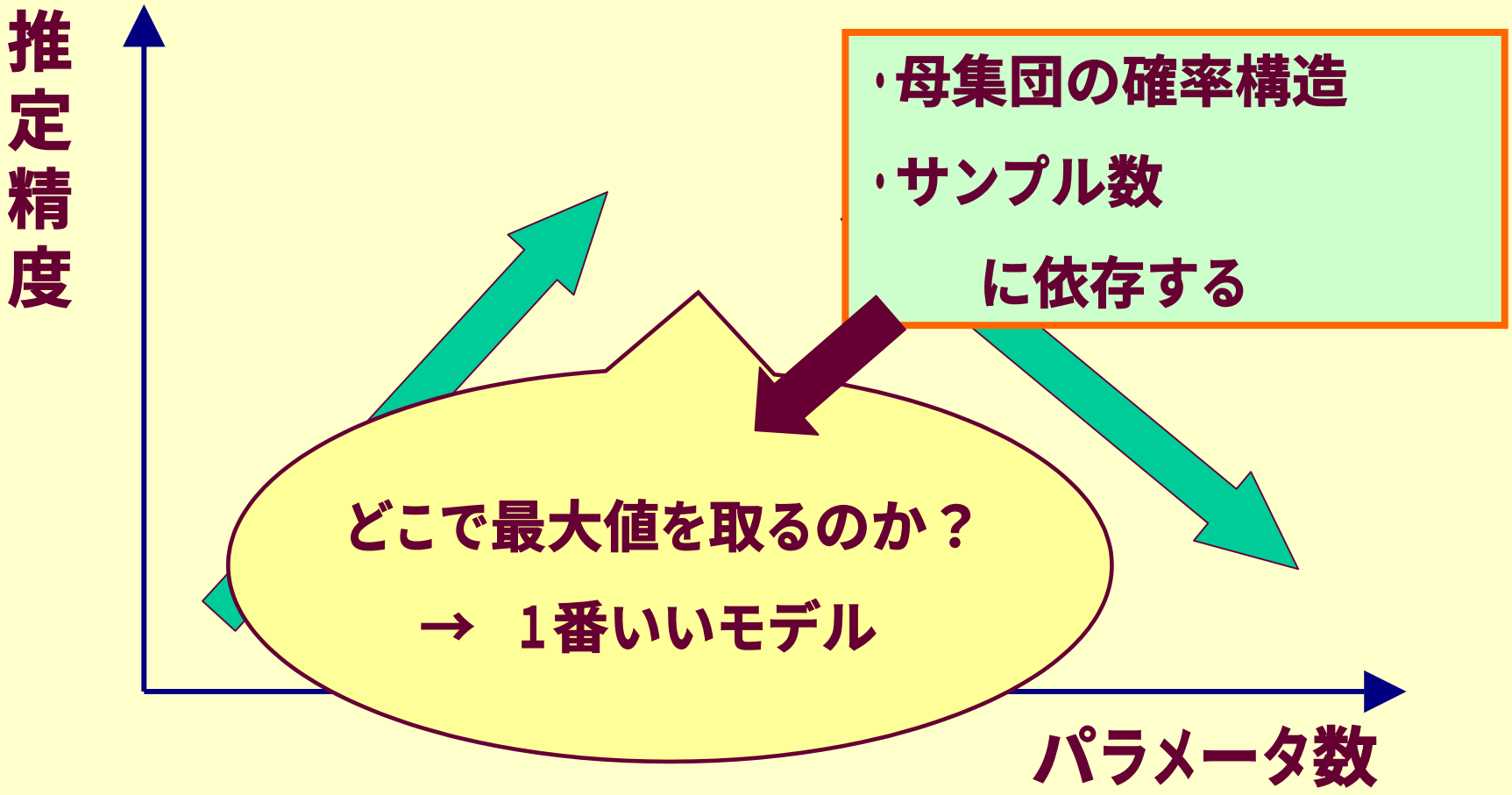
モデル2

$\hat{\mu}_1 = 1000, \quad \hat{\mu}_2 = 1060$

それぞれデータ5個から推定

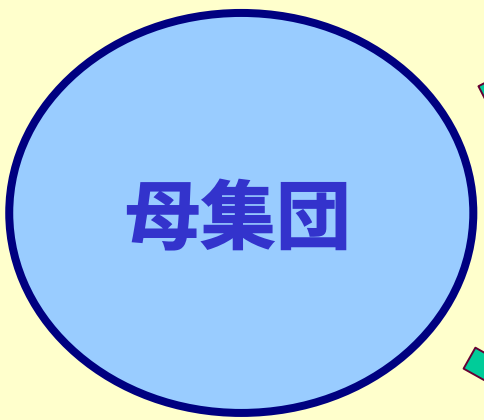
# パラメータ数と推定精度

データがGiven → サンプル数一定 ならば



# 統計的モデル選択の一般論

確率構造を持つ
対象を仮定する




仮説（候補の確率モデル）

モデル1： $p(x \mid m_1, \theta_1)$

モデル2： $p(x \mid m_2, \theta_2)$

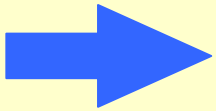
⋮

モデルM： $p(x \mid m_M, \theta_M)$

技術的見地からは適合するモデル

# 現実問題と統計モデル

主に予測やモデル化を目的とするような問題



多変量解析や時系列解析

例えば,,,

- 重回帰モデル
- 判別モデル
- 分類木 (決定木)
- 自己回帰モデル
- 移動平均自己回帰モデル
- ニューラルネットワークモデル

# 重回帰モデル

説明変数: $X_1, X_2, \Lambda\ , X_p$

目的変数: $Y$

} 関係をモデル化

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \Lambda\ + b_p X_p + \varepsilon$$

$\varepsilon$ は

- ①独立性
- ②不偏性
- ③当分散性
- ④正規性

} をみたす誤差（確率変数）

# 判別モデル

**説明変数:** $X_1, X_2, \Lambda\ , X_p$

**目的変数:** $Y$

**(質的変数)**

関係をモデル化

ただし、目的変数が質的変数

$$Z = b_0 + b_1 X_1 + b_2 X_2 + \Lambda\ + b_p X_p$$

$$\begin{cases} Z > 0 \quad \text{ならば} \quad Y = 1 \\ Z \leq 0 \quad \text{ならば} \quad Y = 0 \end{cases}$$

# 分類木

**説明変数:** $X_1, X_2, \Lambda\ , X_p$

**目的変数:** $Y$

説明変数が離散変数のとき

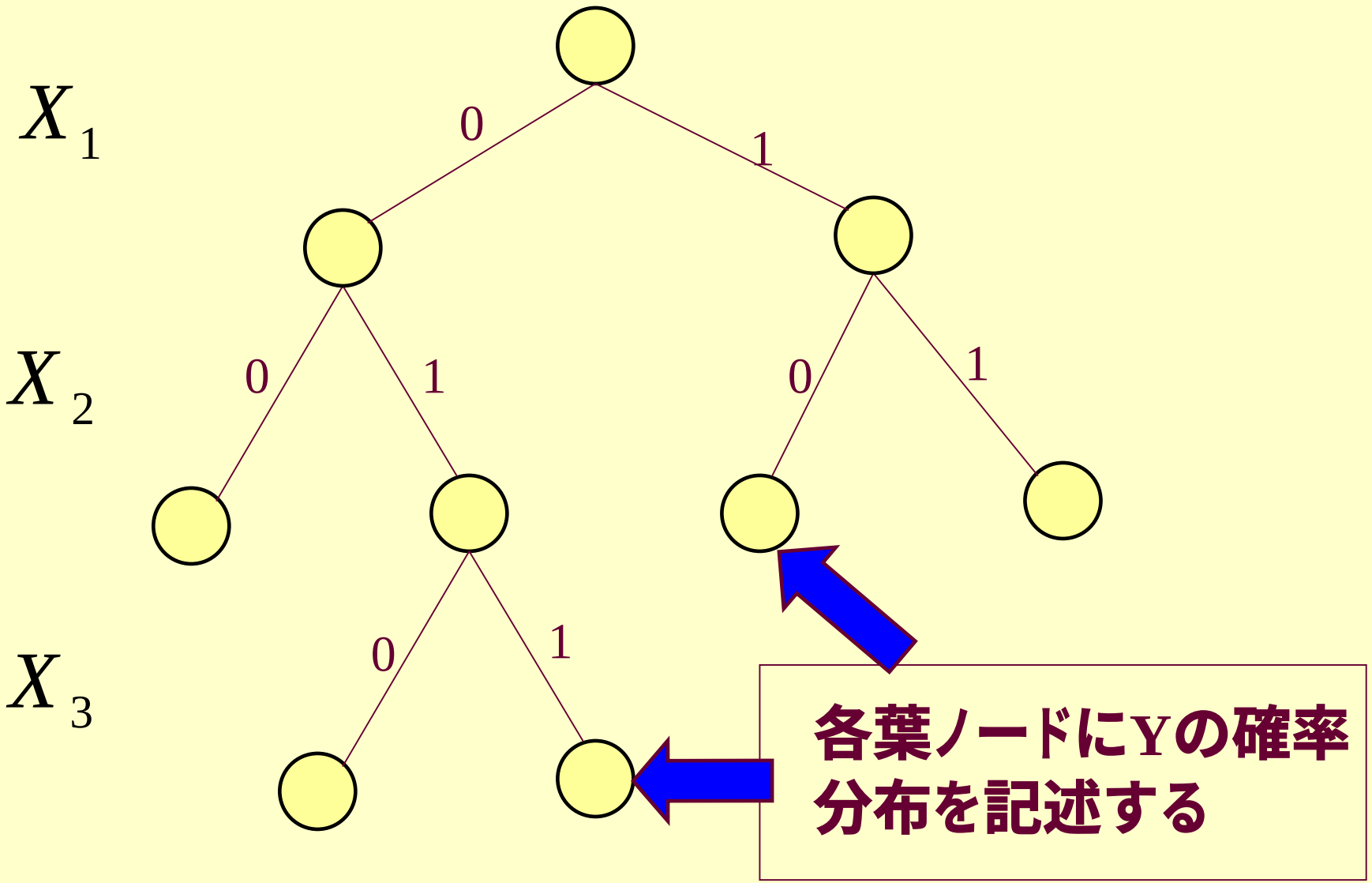
関係をモデル化

**例えば**

$$\begin{cases} X_1, X_2, \Lambda\ , X_p \in \{0,1\} \\ Y \in \{0,1\} \end{cases}$$

0,1は便宜上の値で数値自体に意味はない

# 分類木の例



# ニューラルネット

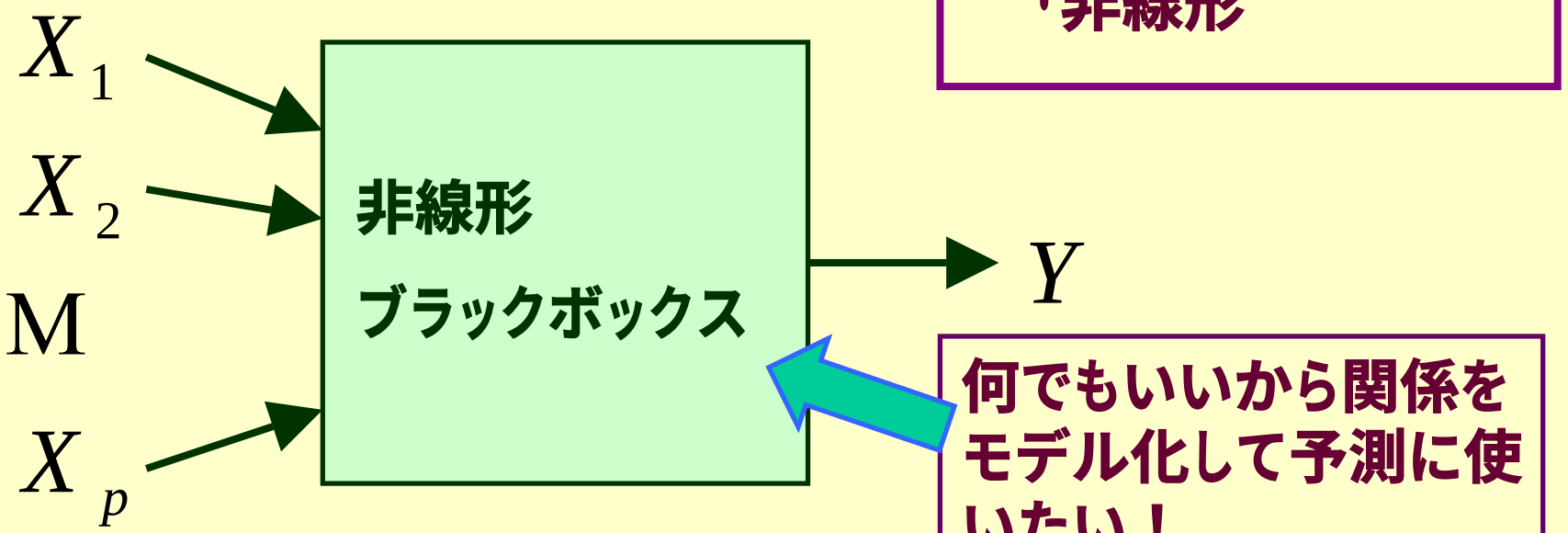
**説明変数:** $X_1, X_2, \Lambda\ , X_p$

**目的変数:** $Y$

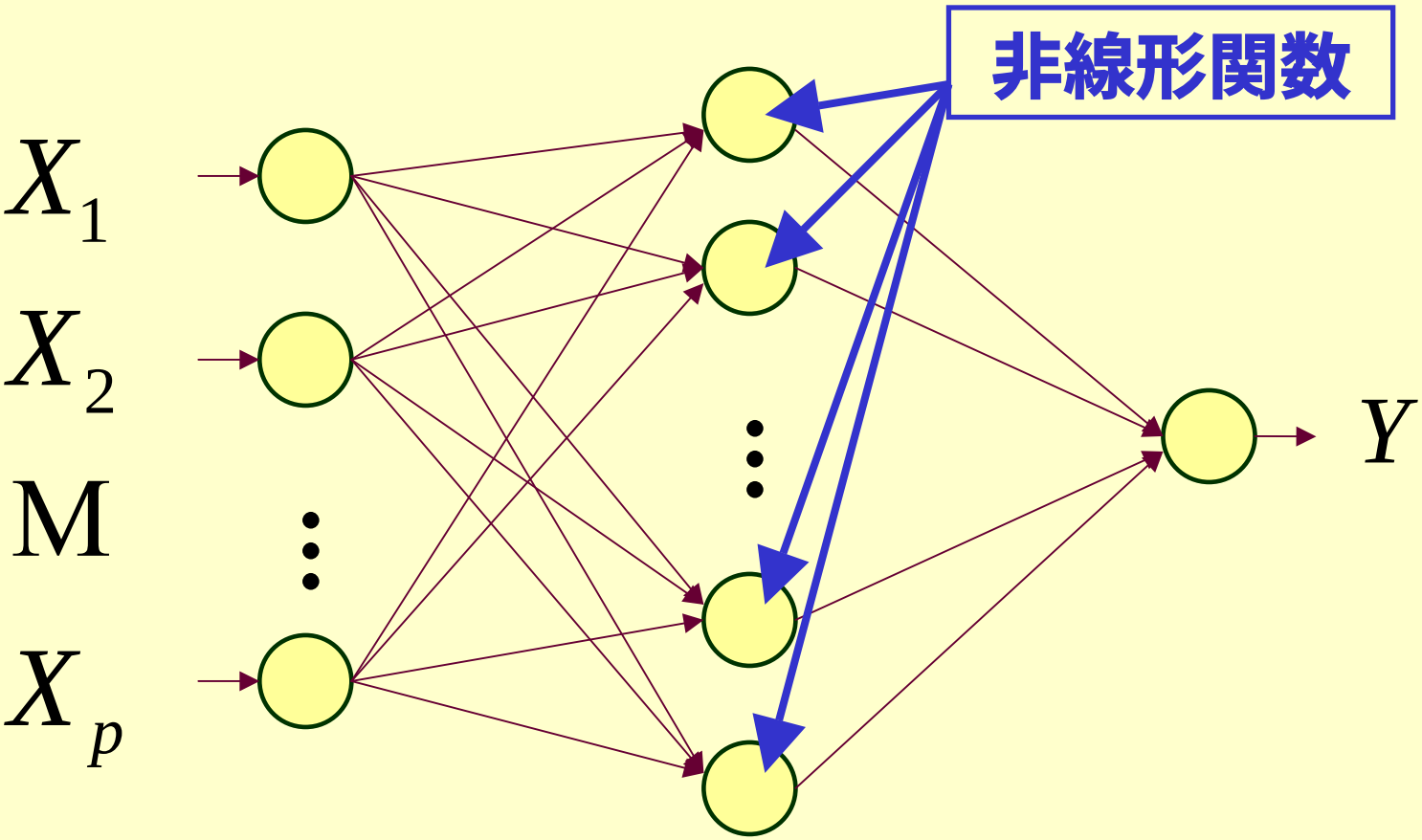
**関係をモデル化**

**ただし、**

- **構造が分からない**
- **非線形**

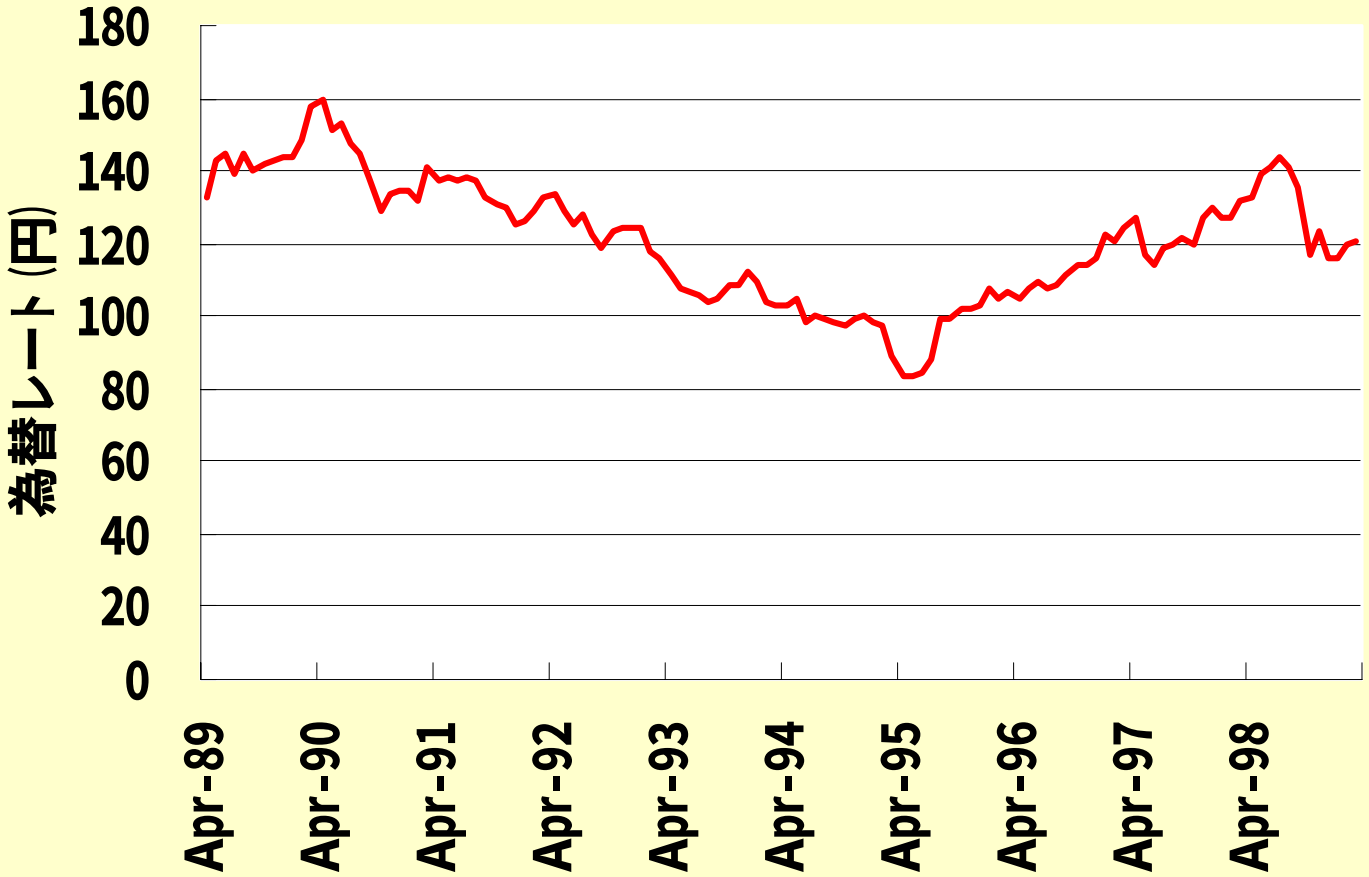


# ニューラルネットモデル



# 時系列モデル

## 時系列データ：為替レートの例




図．89年から98年までの対US$の為替レートの推移

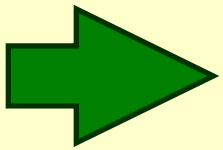
# 自己回帰モデル

$$\Lambda\ \Lambda\ X_{-2}, X_{-1}, X_0, X_1, X_2, \Lambda\ , X_t, \Lambda\ \Lambda$$

時間軸

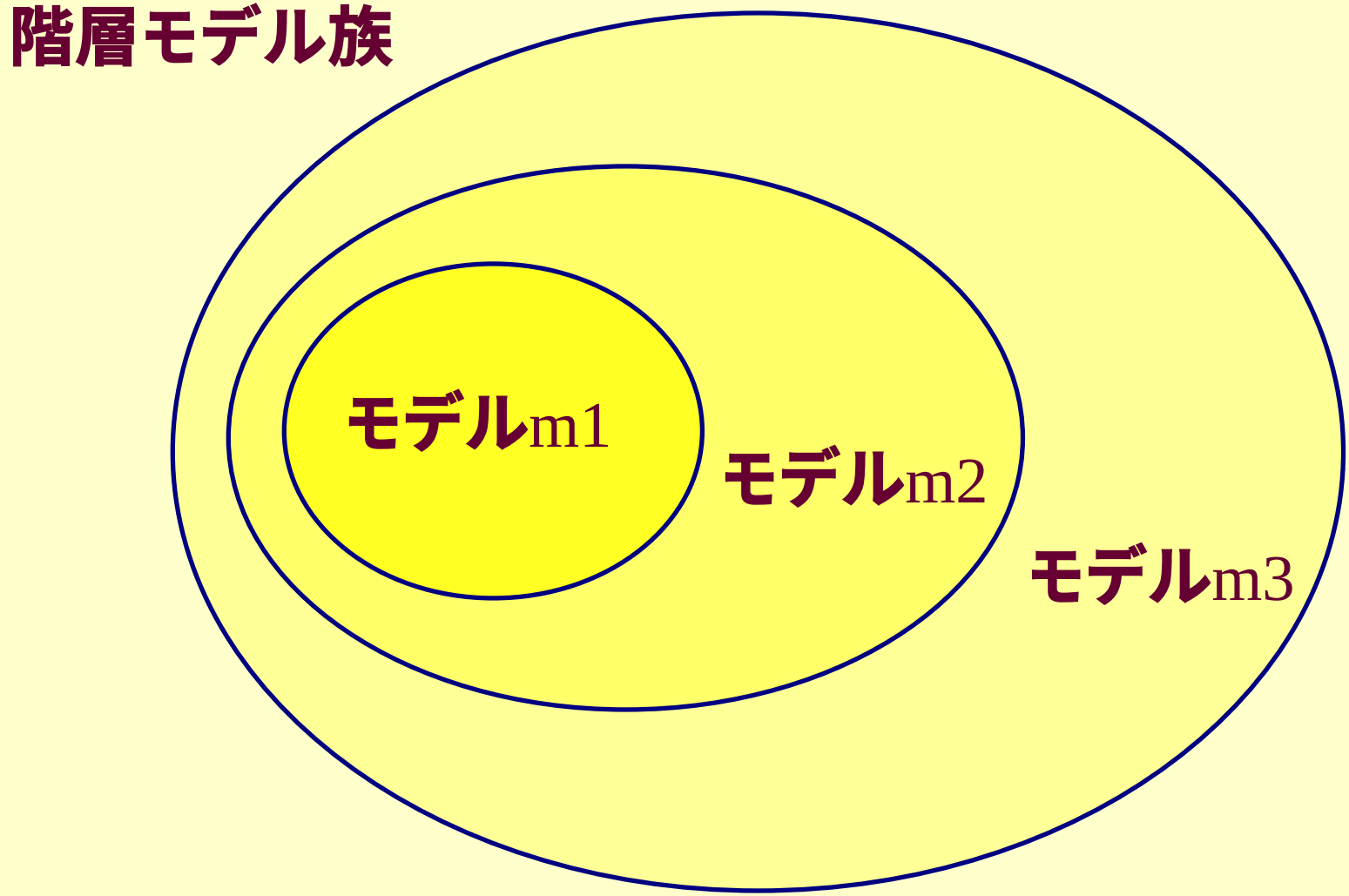
$p$ 次のARモデル

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \Lambda + \alpha_{t-p} X_t + \varepsilon_t$$



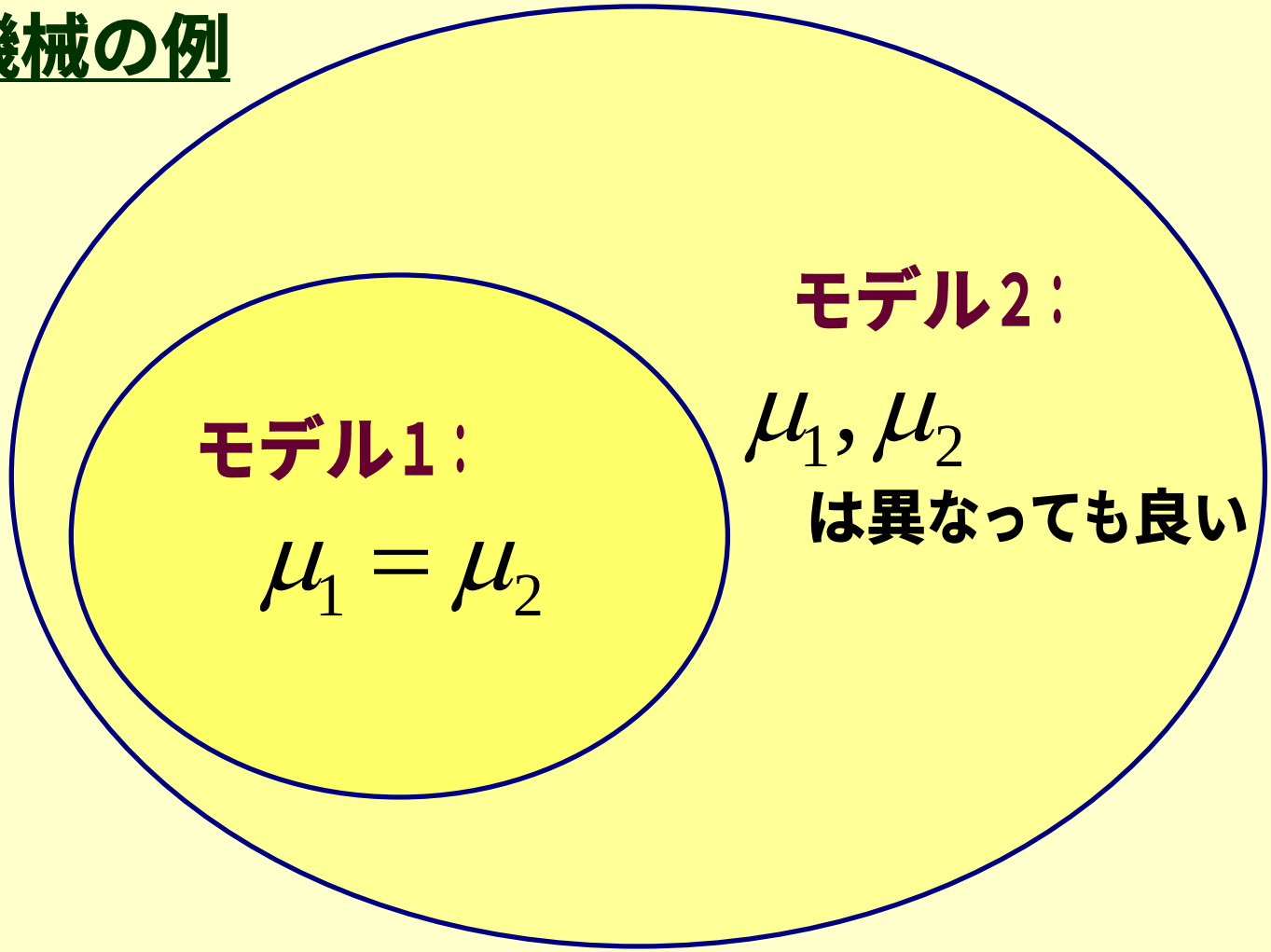
過去の値の線形和で予測

# 一般化：階層モデル族



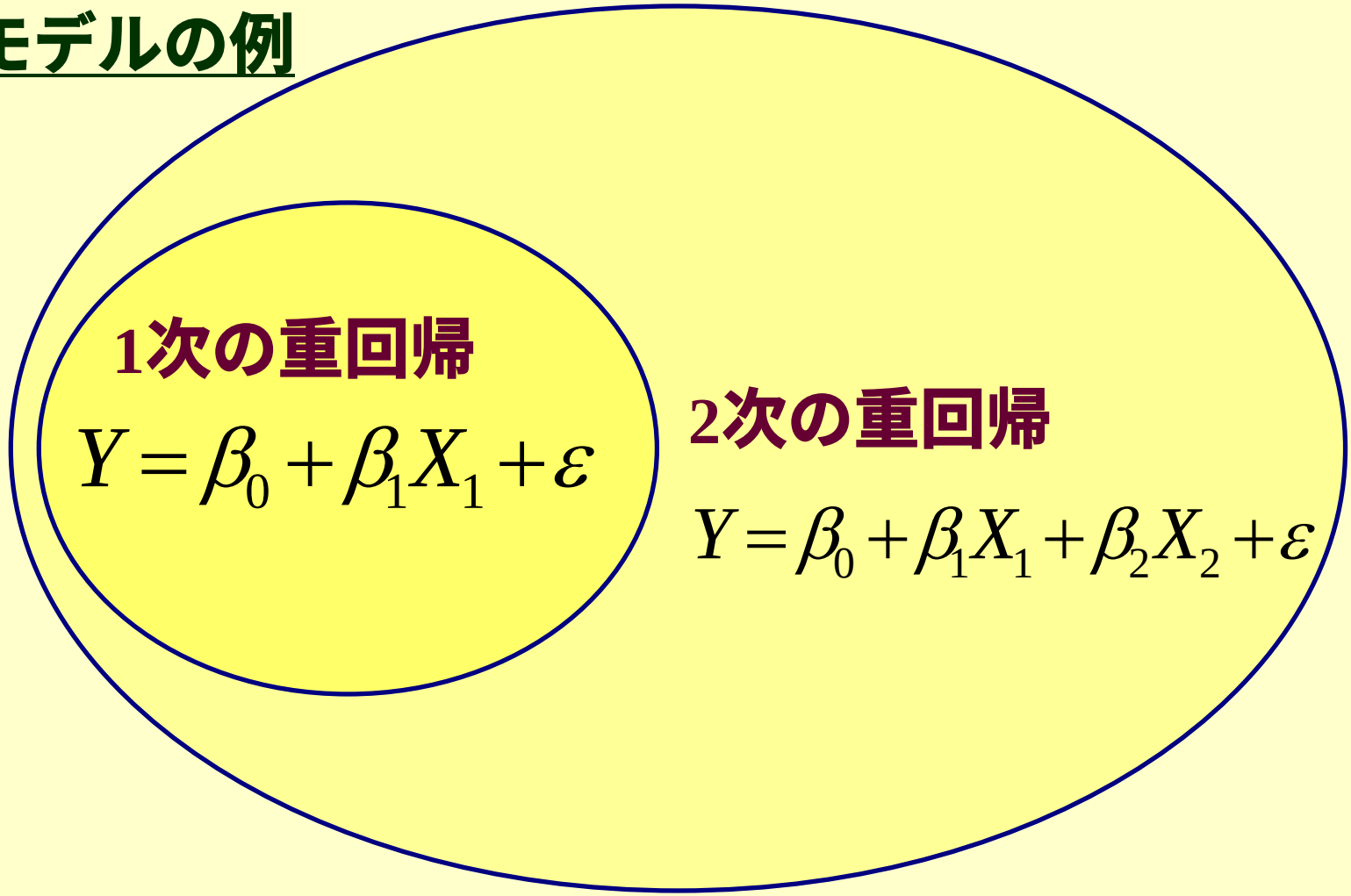
# また簡単な例で(1)

**2台の機械の例**



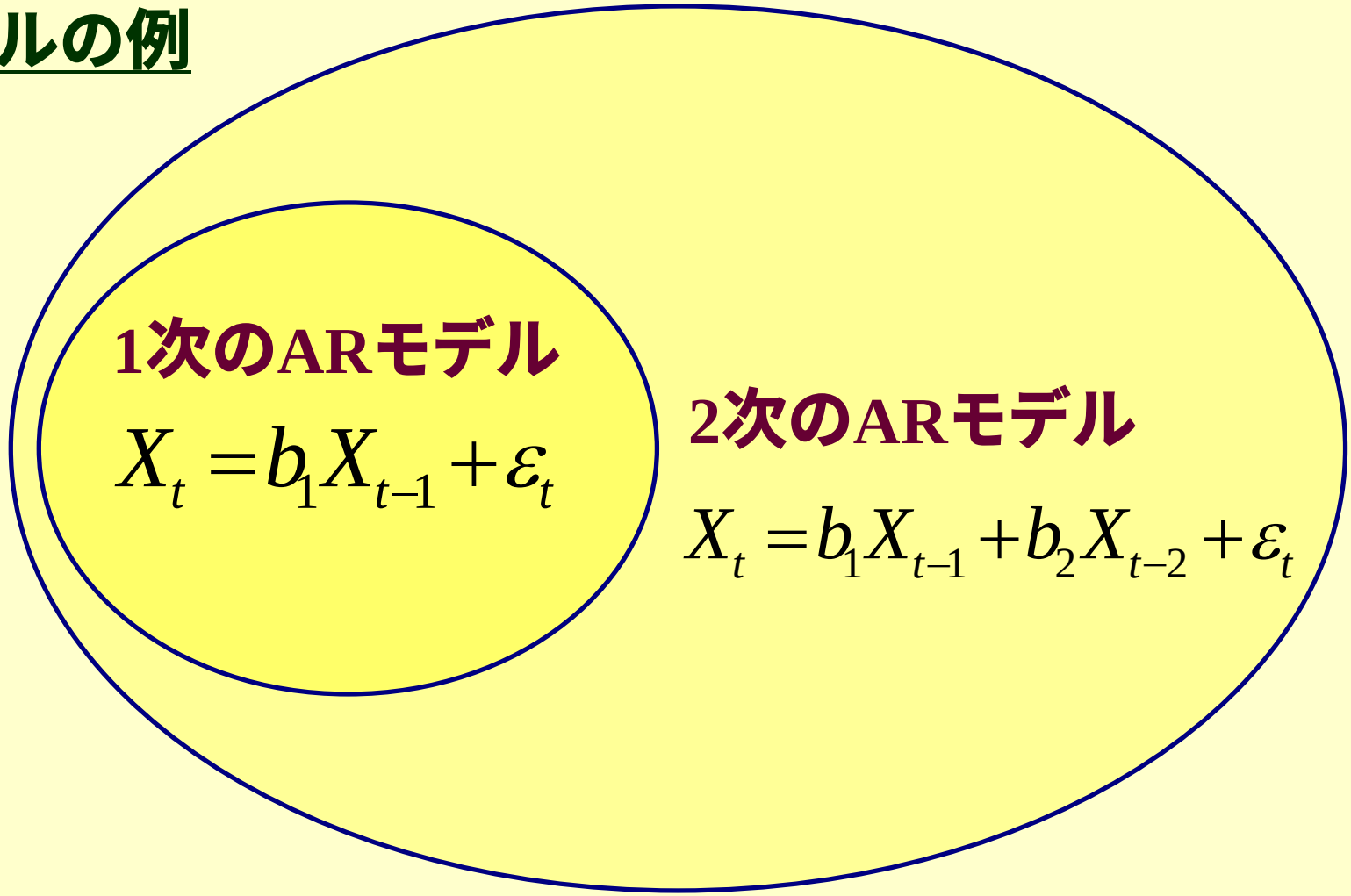
# また簡単な例で(2)

**重回帰モデルの例**



# また簡単な例で（3）

**ARモデルの例**



# 良い統計モデルとは？

- 解析の目的（何のための解析なのか?）を明確にせよ.

データ解析の目的

- 構造解析
- 仮説の検証
- 予測
- ・・・・・・

大まかに分類して

**推定 と 予測**

# 理論的展開

統計的モデル化の目的

↓

目的に見合った評価関数（損失関数）

一般には、真のモデル（仮定）との
何らかの意味での距離

↓

何らかの規準で評価関数を最大にする統計モデルを最良とする

真のモデルが分からないと最小化できない

# 損失関数の分類

## 真のモデルの推定が目的の場合

$L(m, \hat{m})$

真のモデル $m$
推定したモデル $\hat{m}$

の間に

**損失関数**

## 予測が目的の場合

$L(x, \hat{x})$

実現値 $x$
予測値 $\hat{x}$

の間に

**損失関数**

# 損失関数の種類

- 0-1損失 → 離散変数間の問題で一般的
- 絶対誤差損失
- 二乗誤差損失 → 連続変数間の問題で一般的
- 対数損失 → 確率分布間の問題で一般的
- α損失　など

# 統計的モデル選択の基準

- AIC(Akaike Infromation Criteria)
- BIC(Bayesian Information Criteria)
- MDL(Minimum Description Length)
- Cross Validation
- FPE(Final Prediction Error)

**など**

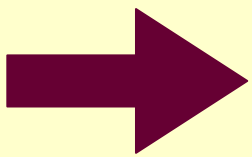
# AIC(Akaike Information Criterion)

- **目的 → 予測**
- **損失関数 → 予測の対数損失関数**

$$AIC(m) = -\log p(x^n \mid m, \hat{\theta}_m) + k_m$$

ただし、$k_m$ : モデル$m$のパラメータ数

$\hat{\theta}_m$: モデル$m$の最尤推定量



**これを最小化するmを最適なモデルとする**

# BIC(Bayesian Information Criterion)

- **目的 → 真のモデルの推定**
- **損失関数 → モデル間の0-1損失関数**

→ **ベイズの事後確率を最大化するモデルの選択**

$$BIC(m) = -\log p(x^n \mid m, \hat{\theta}_m) + \frac{k_m}{2}\log n$$

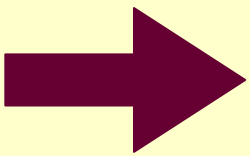
→ **これを最小化するmを最適なモデルとする**

# MDL(Minimum Description Length)

- MDL**原理 → 解析データを最も圧縮できる確率モデルが最良であるとする原理**

**•目的 → データの圧縮**
**•損失関数 → 解析データの対数損失関数**

$$MDL(m) = -\log\, p(x^n \mid m, \hat{\theta}_m) + \frac{k_m}{2}\log n$$



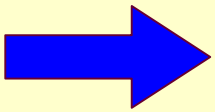
**これを最小化するmを最適なモデルとする**

# Cross Validation

- 目的 → 予測
- 損失関数 → 何でも良い

**データを2つのグループに分割**

一方のデータのみでモデルを推定し、他方のデータを予測する



最も予測精度の高かったモデルを

最良のモデルとする

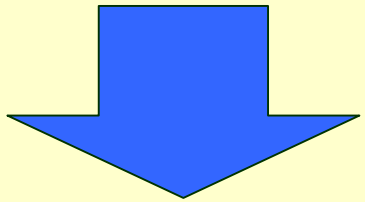
# 具体例

# 実務レベルの注意点

- 規準が同程度のモデルが複数存在したら？
  → 複数のモデルの平均で予測する方法
  → 技術的な考察に基づきモデルの選択する方法　など
- 重回帰モデル等における多重共線性
  → 制御できる変数を取り込む方法　など
- 候補のモデル族が適切でないと意味がない
  → モデル選択規準の前に、候補のモデル族に対して十分な考察を行う必要がある
  → 何のためにデータ解析を行うのか、その目的を明確にする必要がある

# まとめ

- 本編では統計的モデル選択の概略について述べた
- モデル選択は本質的に階層モデル族の入れ子構造に起因する問題である



モデル選択規準は、その目的や背景を理解して用いないと誤った解釈に陥るので注意が必要である