時系列分析と
状態空間モデルの基礎

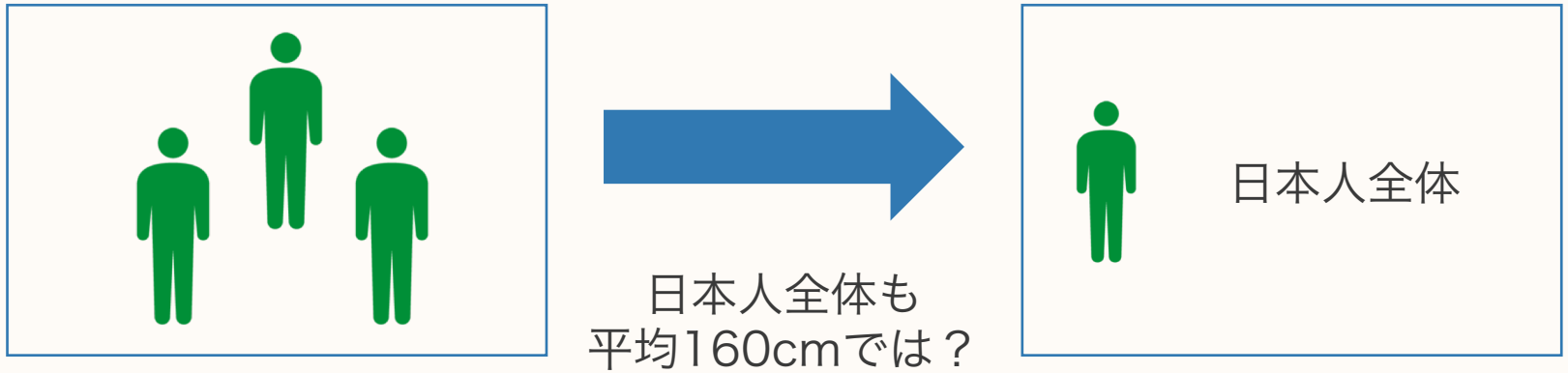
第1部

加藤

基礎・用語の確認

- 標本と母集団

推測統計学の考え方



日本人100人をサンプリング
平均身長160cm

- 標本　：100人の日本人
- 母集団：日本人全体

- 時系列データとは
  - 年月時分秒など一定の間隔でとられた一連のデータ
  - 時系列データ↔トランザクションデータ

用語の確認

- **母数は母集団サイズとは違う (よく間違えてました)**
  - 確率分布を決定する統計量を母数（パラメータ，parameter）と言う
- **サンプル数とサンプルサイズもよく間違える (主に私)**
  - 30 人と 40 人の身長の母平均の差を検定する場合
    - サンプル数：2
    - サンプルサイズ：30と40

母集団と確率分布・標本と確率変数

- サイコロの話
- 確率変数：{ 1,　　2,　　3.　　4,　　5,　　6}
  確率分布：{1/6,　1/6,　1/6,　1/6,　1/6,　1/6}
  - 標本→確率変数のこと
  - 母集団として，ある特定の確率分布を想定する

標本は，母集団の確率分布に従って得られると考える

データ生成過程(DCP)

- 時間変化のある確率分布のこと(確率過程，過程とも呼ぶ)

- 例：2000/1/1と2000/1/2の気温について
- 2000/1/1
  確率変数：{ 1,　2,　3.　4,　5,　6}
  確率分布：{1/6,　1/6,　1/6,　1/6,　1/6,　1/6}
- 2000/1/2
  確率変数：{ 1,　2,　3.　4,　5,　6}
  確率分布：{1/4,　1/6,　1/6,　1/6,　1/6,　1/12}

実際手元にあるのは
それぞれ１つの気温だけ！
↓
なぜ確率分布？

- 「もしも2000/1/1が複数あれば？」→次には異なる気温が得られると想定する
  - １つのデータから理論的な期待値や分散を求める

データ生成過程の構造

- さっきの例だと…

夏は暑くて冬は寒い
昨日が寒かったから今日も寒い
地球温暖化だから年々気温は上がる

- このような過程をモデル化する→時系列モデル
- 時系列モデルのメリット
  - データ生成過程がわかり，理論的な期待値や分散が計算可能
  - 季節の影響やトレンドの有無を判断可能
  - 将来の予測

時系列データの構造

- **自己相関とコレログラム**
  - 昨日の気温が高ければ今日も高い→**正の自己相関**
  - 昨日の気温が高ければ今日は低い→**負の自己相関**
  - １日前と相関があるの？７日前と相関があるの？
    → 自己相関をグラフ化 → **コレログラム**
- **季節成分・周期成分**
  - 年単位の周期性 → 季節成分/季節性

時系列データの構造

- **トレンド**
  - 毎月売り上げが右肩上がり！→ トレンド
  - 例：来月の売り上げ＝今月の売り上げ＋トレンド(例えば２０万円)
- **外因性**
  - RLSキックオフがあるからビッグサイトの利用者が増えた！
  - 自己相関で表せないものを表現
- **ホワイトノイズ**
  - 期待値が0，分散が一定，自己相関が0
  - 「平均0，分散 δ^2 の正規分布」に従うホワイトノイズがよく使われる
    - モデル化が楽だから

時系列データの構造

- トレンド
  - 毎月売
  - 例：来
- 外因性
  - RLSキ
  - 自己相
- ホワイト
  - 期待値
  - 「平均0,
    - モデル化が楽だから

２０万円)

！

よく使われる

時系列データ　＝　短期の自己相関
＋周期的変動
＋トレンド
＋外因性
＋ホワイトノイズ

数式表現の確認

- １～T時点までの時系列データ(y_t はある時点tにおけるデータ)

$$Y_T = \{y_t\}_{t=1}^{T} = \{y_1, y_2, \cdots, y_T\}$$

- t時点のデータが期待値μ，分散　　のδ^2正規分布に従う場合

$$y_t \sim N(\mu, \delta^2)$$

- 期待値・分散

$$\mu_t = E(y_t) \quad \mathrm{Var}(y_t) = E[(y_t - \mu_t)^2]$$

※2000/1/1が無数にあったとしたら，気温は平均的に μ_t

自己共分散

- 一時点前との自己共分散→「1次の自己共分散」

$$\gamma_{1t} = Cov(y_t, y_{t-1}) = E[(y_t - \mu_t)(y_{t-1} - \mu_{t-1})]$$

- K次の自己共分散の場合

$$\gamma_{kt} = Cov(y_t, y_{t-k}) = E[(y_t - \mu_t)(y_{t-k} - \mu_{t-k})]$$

1次の自己共分散がプラス→t時点のデータが期待値より大きかったとしたら,
t-1時点も同じように期待値よりも大きな値になりやすいと期待できる

問題点：自己共分散の最大値最小値はデータによって異なる

自己相関

- 共分散の最小値を−1，最大値を+1に標準化した指標

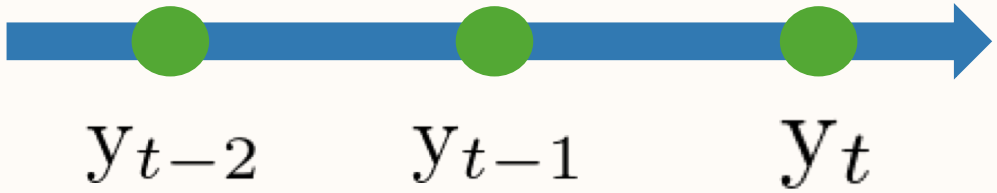
$$\rho_{kt} = Corr(y_t, y_{t-k}) = \frac{Cov(y_t, y_{t-k})}{\sqrt{Var(y_t)Var(y_{t-k})}}$$

- 自己相関が1に近い→「t時点で大きい値ならt-k時点も大きい」
- 自己相関をkの関数とみなしたもの→自己相関関数

偏自己相関

2時点前との自己相関を確認したい！

$y_t = 0.8y_{t-1}$　　$y_{t-1} = 0.8y_{t-2}$



$$y_t = 0.8(0.8y_{t-2})$$

本来は1時点前と関係があるだけ

But

2時点前と関係があるように見える

1時点前との関係性を排除した上で,

2時点前との自己相関を計算したい

偏自己相関

偏自己相関

$$\hat{y}_t = \alpha y_{t-1}$$

1時点前と現在時点との関係式(ハットは推定量)

係数αは $\mathrm{E}[(y_t - \hat{y}_t)^2]$ を最小とする

$$\hat{y}_{t-2} = \beta y_{t-1}$$

2時点前も同様に表現

1時点前のデータから表現する
ことができなかった残り

こちらも(現在から見て)
1時点前から２地点前を
表現できなかった残り

$$\mathrm{P}_{t2} = \frac{Cov(y_t - \hat{y}_t, y_{t-2} - \hat{y}_{t-2})}{\sqrt{Var(y_t - \hat{y}_t)Var(y_{t-2} - \hat{y}_{t-2})}}$$

1時点前でのデータでは表現できなかった残り同士で相関を取る！

ホワイトノイズ

ε_t ：t時点のホワイトノイズ

$$\mathrm{E}[\varepsilon_t] = 0$$

$$\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-k}) = \begin{cases} \delta^2, k = 0 \\ 0, k \neq 0 \end{cases}$$

要は

- 期待値が0
- 分散が一定
- (同時刻以外の)自己相関が0
  (k=0の共分散は分散と等しい)

未来を予測する情報がほとんど含まれていない
純粋な雑音

$\varepsilon_t \sim W.N.(\delta^2)$

ε_tがホワイトノイズに従う

$\varepsilon_t \sim N(0, \delta^2)$

平均0，分散δ^2の正規分布に従うホワイトノイズ

iid系列・ランダムウォーク・確率的トレンド

- Independent and Identically Distributed
  - データが独立であること
  - 前ページの正規分布に従うホワイトノイズはiid系列

$$y_t \sim iid(\mu, \delta^2)$$

- ランダムウォークはiid系列の累積和からなる確率過程
- 例：$y_t = y_{t-1} + \varepsilon_t,\ \varepsilon \sim N(0, \delta^2)$

$$y_t = \delta + y_{t-1} + \varepsilon_t,\ \varepsilon \sim N(0, \delta^2)$$

ドリフト率
線形トレンドを表現
(毎月売り上げが20万増える的な)

線形トレンドの増減率が確率的に
変化している！
↓
ランダムウォーク＝確率的トレンド