2020年度後期
応用確率統計

統計的推論

阪口 啓
sakaguchi@mobile.ee.titech.ac.jp

2020年10月30日

講義スケジュール後半

	日付	教科書	内容	演習
第8回	10月30日	9.1, 9.2	統計的推論（推測）	なし
第9回	11月6日	9.3, 10.1	平均と分散の不偏推定	あり
第10回	11月10日	11, 13	最尤推定とベイス推論	あり
第11回	11月13日	12	仮説検定、分類、機械学習	あり
第12回	11月17日	配布資料	回帰、予測、機械学習	あり
第13回	11月20日	配布資料	確率過程、相関、予測	あり
第14回	11月24日		理解度確認総合演習	

確率・統計とは？

• ランダムな（確率的な）ものには情報がある

• ランダムな観測値から情報を統計的に取り出す

■ 確率論（基礎的）

• 確率的現象を理論的に扱う

• 確率空間を設定しその上で理論体系を構築

■ 統計学（応用的）

• 現実に起こる現象を実験的に処理する

• 確率的構造を用いてその背後にある真実を推測する

⇩

統計的推論 (Statistical inference)

統計的推論の分類

- 母数の推定 (Parameter estimation)
  - 観測データから未知母数(統計パラメタ)を推定する方法
  - 例:標本平均、中央値、標本分散
- 仮説検定、分類 (Hypothesis test, Classification)
  - 観測データと想定するデータが一致するかどうかを確認する方法
  - 例:PCR検査、顔認証、音声認識
- 回帰、予測 (Regression, Prediction)
  - 観測データから2つ以上の物理量の関係を表現する方法
  - 例:電圧と電流の関係、GPAと採用確率、天候と収穫量
- 確率過程、予測 (Stochastic process, Forecast)
  - 未知母数が時間の関数である観測データから未来を予測する方法
  - 例:位置推定、コロナ感染者数、AlphaGo

母数の推定とは？

■ 観測データ

例:　$X_i = \theta + \varepsilon_i \qquad (i = 1, \cdots, n)$

確率変数　未知母数　誤差・偶然変動

■ 母数の推定

観測値 $X_1, \cdots, X_n$ から未知母数 θ を推定すること

■ 仮定

1. $\varepsilon_1, \cdots, \varepsilon_n$ は互いに独立
2. $\varepsilon_1, \cdots, \varepsilon_n$ は同じ分布

i.i.d.
(independent identically distributed)

3. 誤差の平均 $E(\varepsilon_i) = 0$
4. 誤差の分散 $V(\varepsilon_i) < \infty$

推定量と推定値の定義

■ 推定量 （Estimator)

確率変数 $X_1, \cdots, X_n$ から未知母数 θ を推定する方式

$$\hat{\theta}(X_1, \cdots, X_n) \longleftarrow \text{関数、確率変数}$$

■ 推定値 （Estimate)

実現値 $x_1, \cdots, x_n$ を推定量 $\hat{\theta}$ に代入して得られる値

$$\hat{\theta}(x_1, \cdots, x_n) \longleftarrow \text{関数の値、実現値}$$

平均の推定量の例

■ 標本平均 (Sample average)

$$\hat{\theta} = \frac{X_1 + \cdots + X_n}{n}$$

■ 中央値 (Median)

大きさの順に並べ替え

$$\{\tilde{X}_1, \cdots, \tilde{X}_n\} = \text{sort}\{X_1, \cdots, X_n\}$$

$$\hat{\theta} = \begin{cases} X_{(n+1)/2} & n : \text{odd} \\ \left(X_{n/2} + X_{n/2+1}\right)/2 & n : \text{even} \end{cases}$$

■ 加重平均 (Weighted average)

$$\sum_{i=1}^{n} c_n = 1$$ ← データの重み

$$\hat{\theta} = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$$

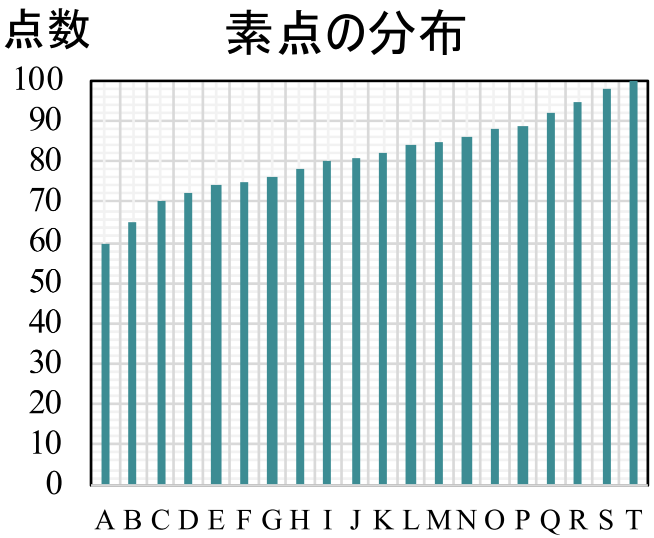
■ 幾何平均 (Geometric average)

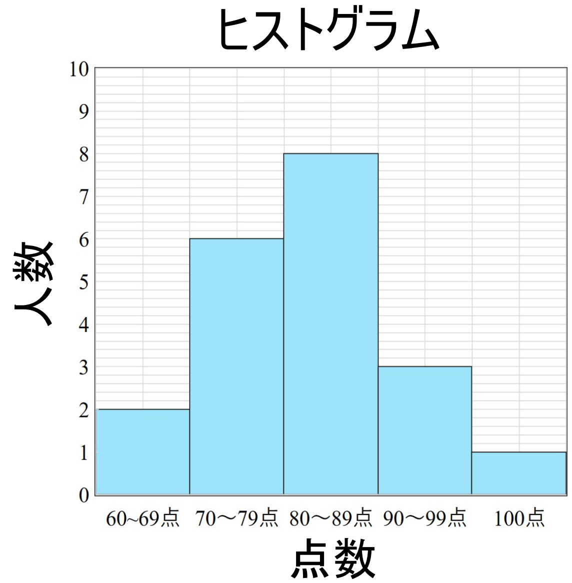
$$\hat{\theta} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$

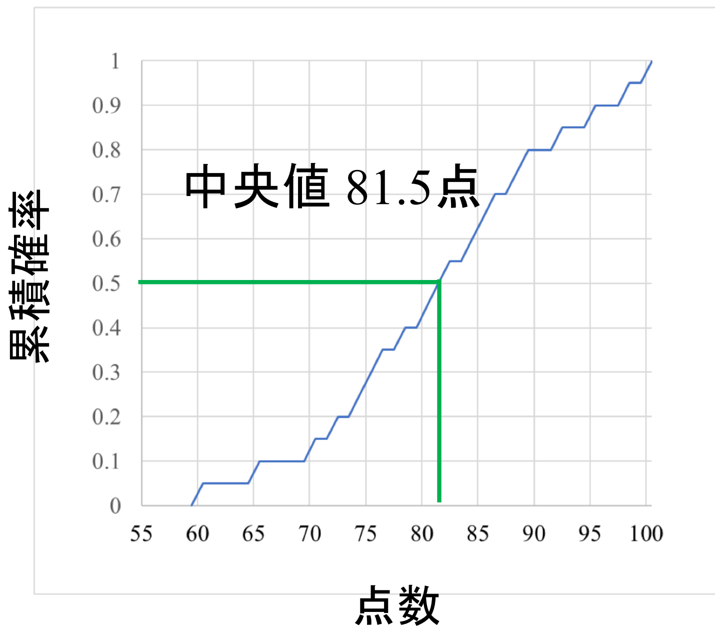
X_i は正の数

テストの点数の例

	点数
A君	60
B君	65
C君	70
D君	72
E君	74
F君	75
G君	76
H君	78
I君	80
J君	81
K君	82
L君	84
M君	85
N君	86
O君	88
P君	89
Q君	92
R君	95
S君	98
T君	100







標本平均

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = 81.5$$

最尤推定

■ 最尤推定 (Maximum likelihood estimation)

誤差の分布 $f(\varepsilon_i)$ が既知のときに観測データから未知母数を推定する方法

■ 観測値の同時確率密度関数

$$f(X_1, X_2, \cdots X_n, \theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

確率変数 X_i の確率密度関数 $P(X_i | \theta)$ を θ の関数として $f(X_i, \theta)$ と表記

■ 尤度関数

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

■ 最尤推定量

$$\hat{\theta}^* = \arg\max_{\theta} L(\theta)$$

ベイズ学習

■ ベイズ学習 (Bayesian estimation)

未知母数の事前確率分布 $P(\theta)$ を想定し
観測データから事後分布を学習する方法

■ ベイズの定理

$$P(\theta, X) = P(X|\theta)P(\theta) = P(\theta|X)P(X)$$

■ ベイズ推定量

尤度関数 事前確率

$$P(\theta|X_1,\cdots,X_n) = \frac{P(X_1,\cdots,X_n|\theta)P(\theta)}{P(X_1,\cdots,X_n)}$$

事後確率

統計的検定と分類とは？

■ 統計的検定

観測されたデータと想定する仮定が一致するかどうかを確認する方法

■ 例

- クラスAのテストの平均点とクラスBのテストの平均点は一致しているか否か？
- PCR検査の結果は陽性か否か？

■ 統計的分類

観測されたデータに最も近しい分類候補を選択するする方法

■ 例

- 顔認証の結果はAさんかBさんか？
- 音声認識の結果はAと言っているかBと言っているか？

仮説検定と分類の枠組み

■ 観測データ

例: $X_i = \theta + \varepsilon_i \quad (i = 1, \cdots, n)$

■ 仮説検定

帰無仮説 H_0:「 $\theta = \theta_0$ 」

対立仮説 H_1:「 $\theta \neq \theta_0$ 」

⇒

有限の観測データを用いた判定法

仮説検定

■ 分類

分類 Class #0:「 $\theta = \theta_0$ 」

分類 Class #1:「 $\theta = \theta_1$ 」

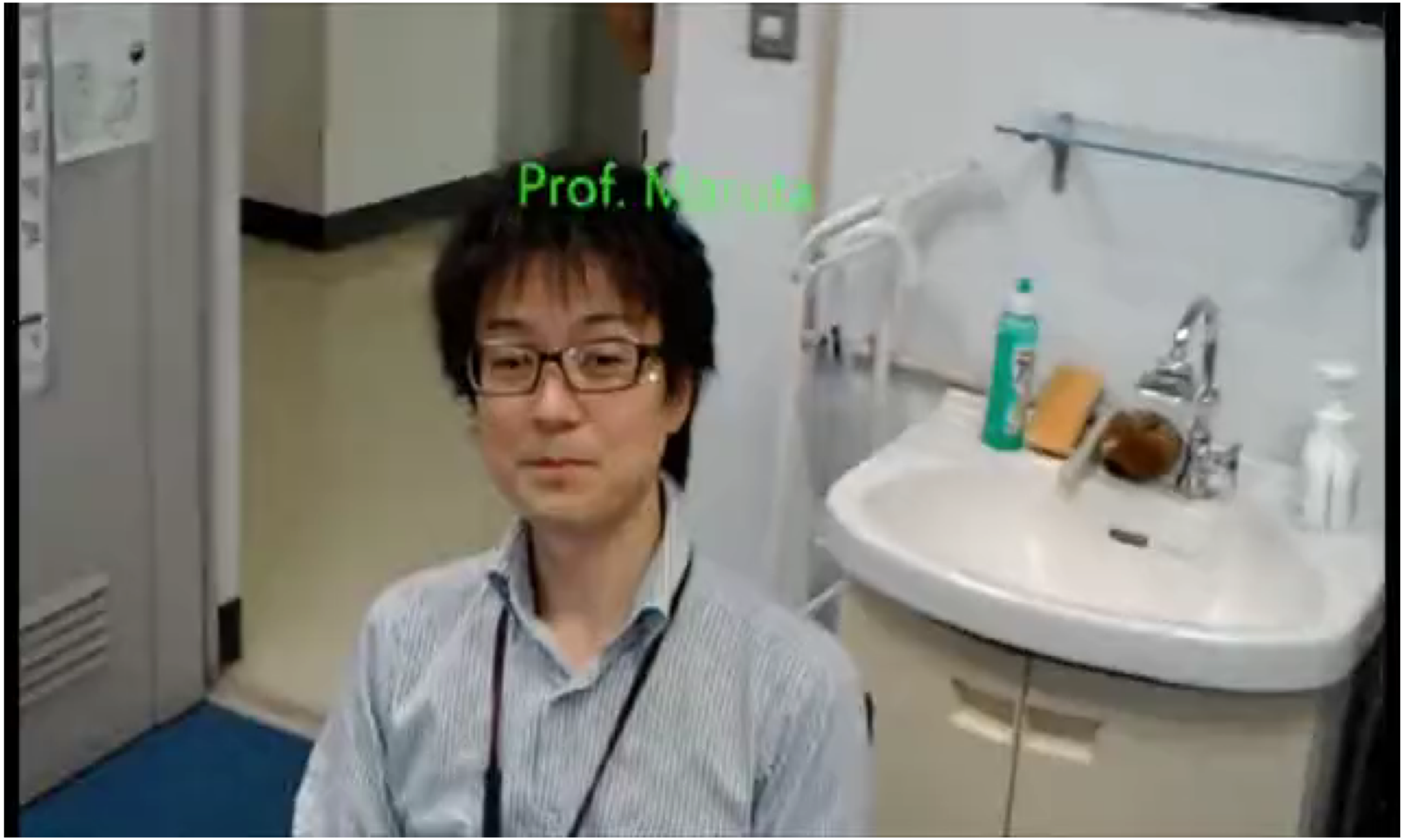
⇒

有限の観測データを用いた判定法

分類(機械学習)

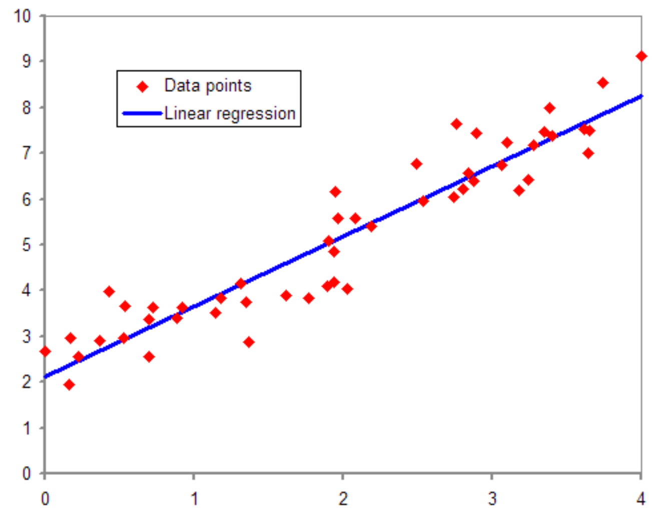
顔認証の例

θ_0 = Prof. Maruta, θ_1 = Mr. Yu, θ_2 = Prof. Sakaguchi



線形回帰と予測とは？

■ 観測データ（2つの物理量の関係）



従属な確率変数　独立変数

例：　$Y_i = ax_i + b + \varepsilon_i \quad (i = 1, \cdots, n)$

未知母数（回帰係数）　誤差・偶然変動

■ 回帰係数の推定法

誤差の分布未知のとき ⟶ 最小二乗法（分散最小）

誤差の分布既知のとき ⟶ 最尤法（漸近不偏、分散最小）

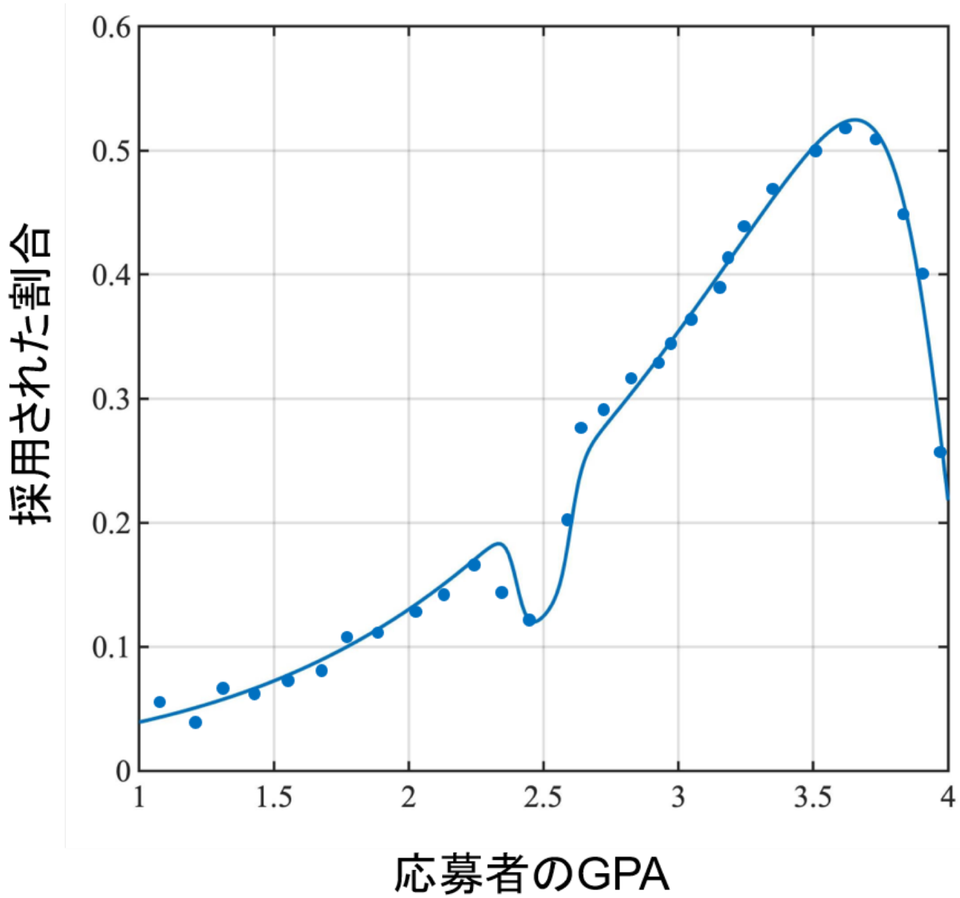
回帰係数の確率推定 ⟶ ベイス学習（事後確率分布の推定）

■ 予測（内挿、外挿）

新規観測値 x_* ⟶ 出力値を推定 y_*

回帰分析の例

$$y_* = g_1(x_*) + g_2(x_*) - 1 + 0.1(g_3(x_*) + g_4(x_*) - 1)$$



$$g_1(x) = \frac{1}{1+e^{-(-4.5+1.3x)}}$$

$$g_2(x) = \frac{1}{1+e^{-(35.4+8.8x)}}$$

$$g_3(x) = \frac{1}{1+e^{-(110.4-46x)}}$$

$$g_4(x) = \frac{1}{1+e^{-(-119.6+46x)}}$$

確率過程と予測とは？

■　多変量統計学（解析）

未知母数は *X* の関数

観測データ:　$Y_i(X) = \theta(X) + \varepsilon_i \quad (i = 1, \cdots, n)$

ε_i は i.i.d. で正規分布 $N(0, \sigma^2)$ に従う

２つの確率変数（変量）*X* と *Y* の関係を統計的に解析

■　確率過程

X が時間　t　のとき $Y_i(t)$ を確率過程と呼ぶ

例:　位置、株価、会話、雑音、コロナ感染者数

■　予測

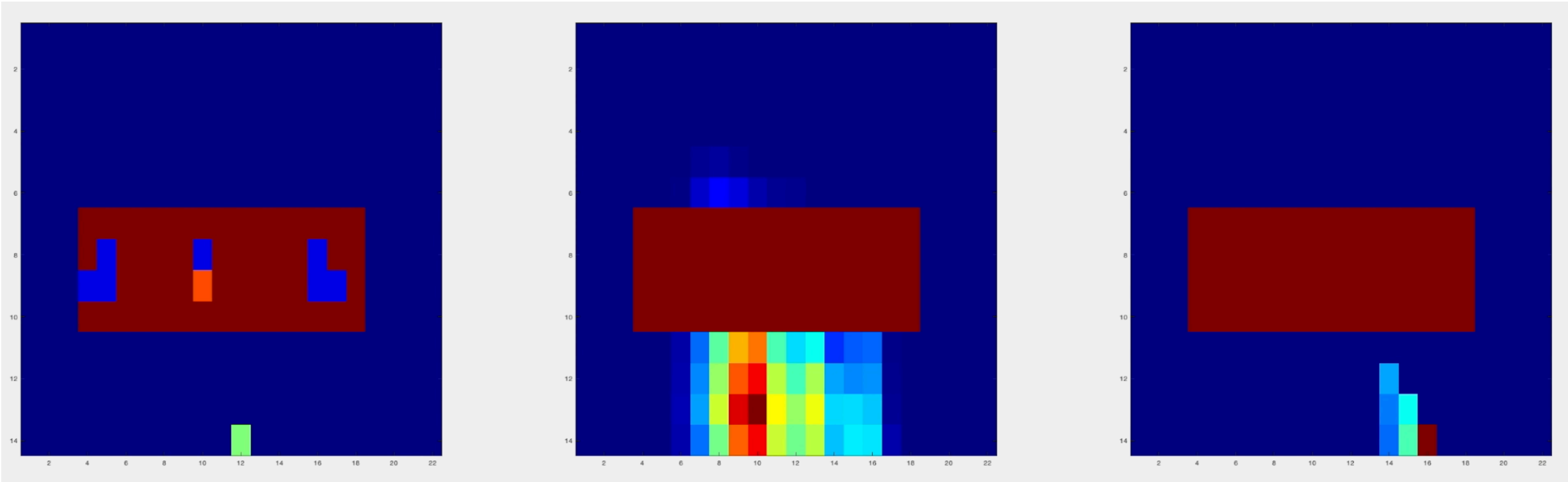
過去の推定値 $\hat{\theta}(t-1)$ と観測データ $Y_i(t)$ から次の $\hat{\theta}(t)$ を予測

人感センサを用いた位置推定の例

$\hat{\theta}(t) = u(t)$: 時刻 t における（人の）位置

人の位置と人感センサ　人の位置の遷移確率　人の位置の事後確率

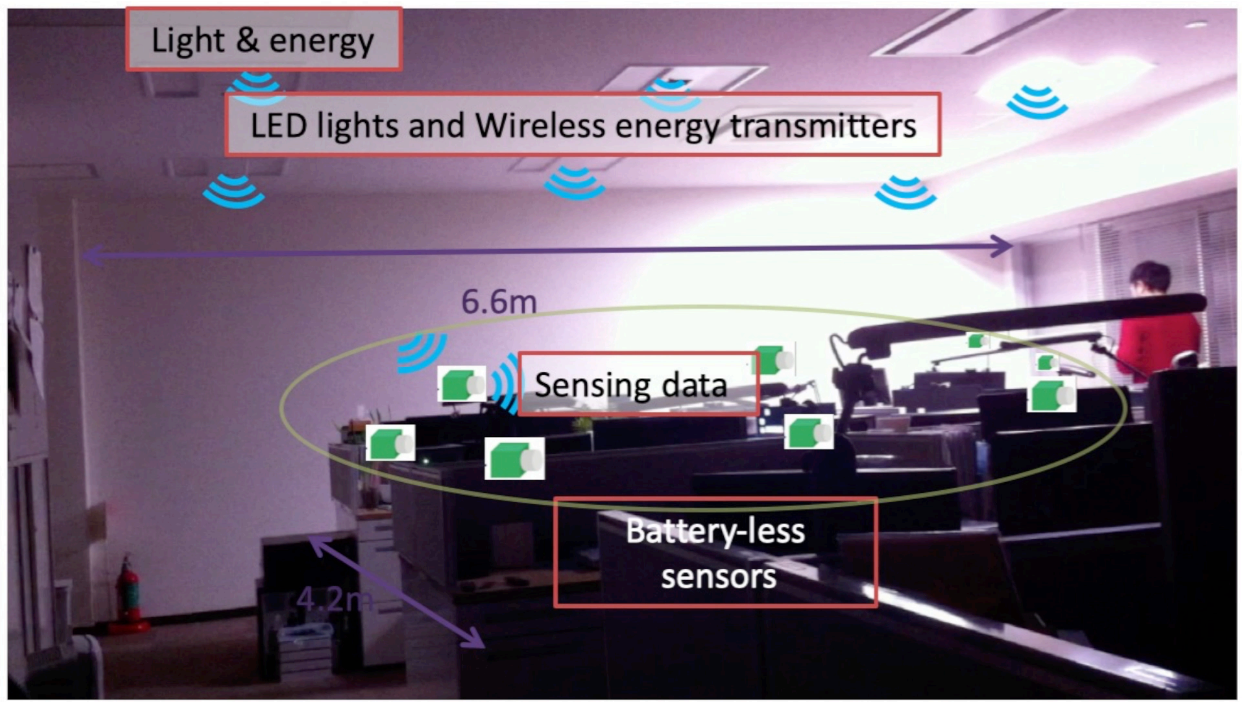
$P(\mathbf{F}_t|\mathbf{u}_t)$　$P(\mathbf{u}_t) = \sum_{\mathbf{u}_{t-1}} P(\mathbf{u}_t|\mathbf{u}_{t-1})P(\mathbf{u}_{t-1})$　$P(\mathbf{u}_t|\mathbf{F}_t) = P(\mathbf{F}_t|\mathbf{u}_t)\,P(\mathbf{u}_t)$



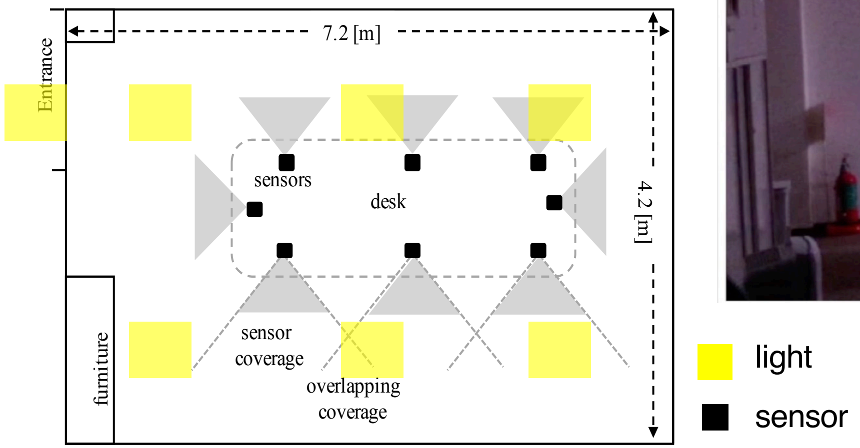
位置推定による照明制御の実験

バッテリーレス人感センサを用いた人の位置推定とその結果を用いた照明の制御

実験結果



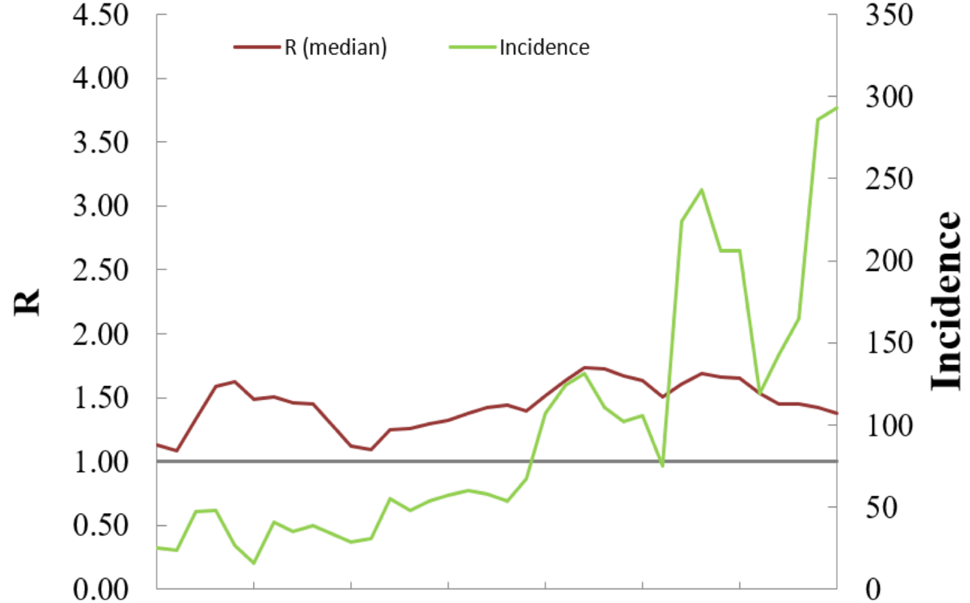
実験環境



コロナ感染者数の例

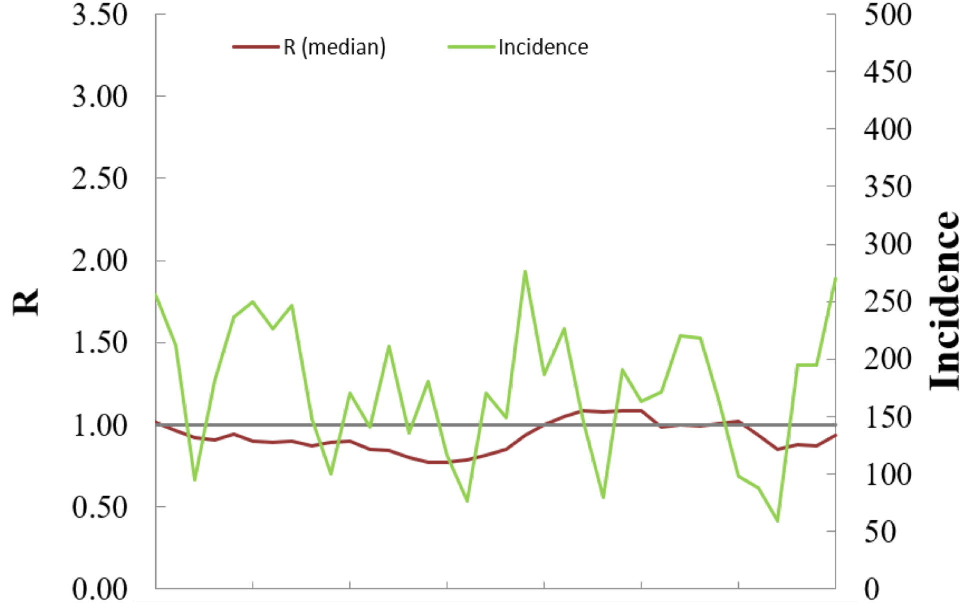
$\hat{\theta}(t) = R(t)$: 実効再生産数（1人から何人に感染するかを表す指標）

新型コロナウィルス
東京都感染者数推移



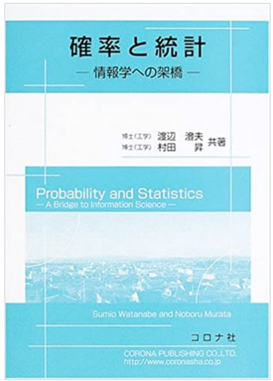
6月12日～7月19日

新型コロナウィルス
東京都感染者数推移



8月23日～9月27日

教科書・参考書



■ 本講義の教科書

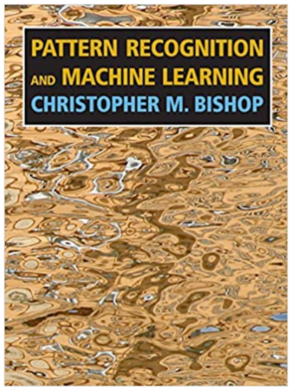
渡辺 澄夫, 村田 昇 共著
“確率と統計
–情報学への架け橋–”
コロナ社, 2017.

■ 統計学の重要性の理解

西内 啓 著
“統計学が最強の学問である”
ダイヤモンド社, 2013.

■ 機械学習の数学的背景

須山 敦志 著, 杉山 将 監修
“ベイズ推論による機械学習”
講談社, 2017.



■ 機械学習のバイブル

C.M. Bishop,
“Pattern Recognition and
Machine Learning”
Springer, 2006.

まとめ

- 母数の推定 (Parameter estimation)
  - 観測データから未知母数（統計パラメタ）を推定する方法
  - 母数の推定: $X_i = \theta + \varepsilon_i \ \ (i = 1, \cdots, n) \ \ \rightarrow \ \ \hat{\theta}$

- 仮説検定、分類 (Hypothesis test, Classification)
  - 観測データと想定するデータが一致するかどうかを確認する方法
  - 仮説検定:「 $\theta = \theta_0$ or $\theta \neq \theta_0$ 」、分類:「 $\theta = \theta_0$ or $\theta = \theta_1$ 」

- 回帰、予測 (Regression, Prediction)
  - 観測データから2つ以上の物理量の関係を表現する方法
  - 線形回帰: $Y_i = a x_i + b + \varepsilon_i \ \ (i = 1, \cdots, n) \ \ \rightarrow \ \ \hat{a}, \hat{b}$

- 確率過程、予測 (Stochastic process, Forecast)
  - 未知母数が時間の関数である観測データから未来を予測する方法
  - 予測: $Y_i(t) = \theta(t) + \varepsilon_i \ \ (i = 1, \cdots, n) \ \ \rightarrow \ \ \hat{\theta}(t+1)$