

統計的モデリング基礎⑬

～因果推論（準実験）～

鹿島久嗣
(情報学科 計算機科学コース)

今回の話題： 因果推論

- 準実験：データから因果を導くにはどうしたらよいか
 - 回帰不連続デザイン
 - 層別解析／回帰モデル
 - マッチング／傾向スコア
 - 差の差（DID）法

ランダム化試験（RCT）： 因果関係を導く方法

- ランダム化試験（RCT）：
 - 因果関係を導くことができる
 - 介入群と対照群を「ランダムに」割り付け、結果を比較する
 - 原因変数に影響を受けない変数の分布が同じになる
- いつでもRCTができるわけではない：
 - そもそもできない（倫理的にできないなど）
 - できたとしても完全にランダムな割り付けを実行できない
 - ◆ 案内を出しても実行しない（ノンコンプライアンス）など

準実験： データから因果関係を導く方法

- 準実験：すでにあるデータから因果関係を導きたい
 - あたかもランダムな割り付けが行われたかのような状況を作り出す
 - 交絡因子を固定する
 - 反事実を疑似的に作り出す
- 準実験の方法：
 - 回帰不連続デザイン
 - 層別解析／回帰モデル
 - マッチング／傾向スコア
 - 差の差（DID）法

回帰不連続デザインと層別分析

回帰不連続デザイン： 割り付けの境界線に注目する

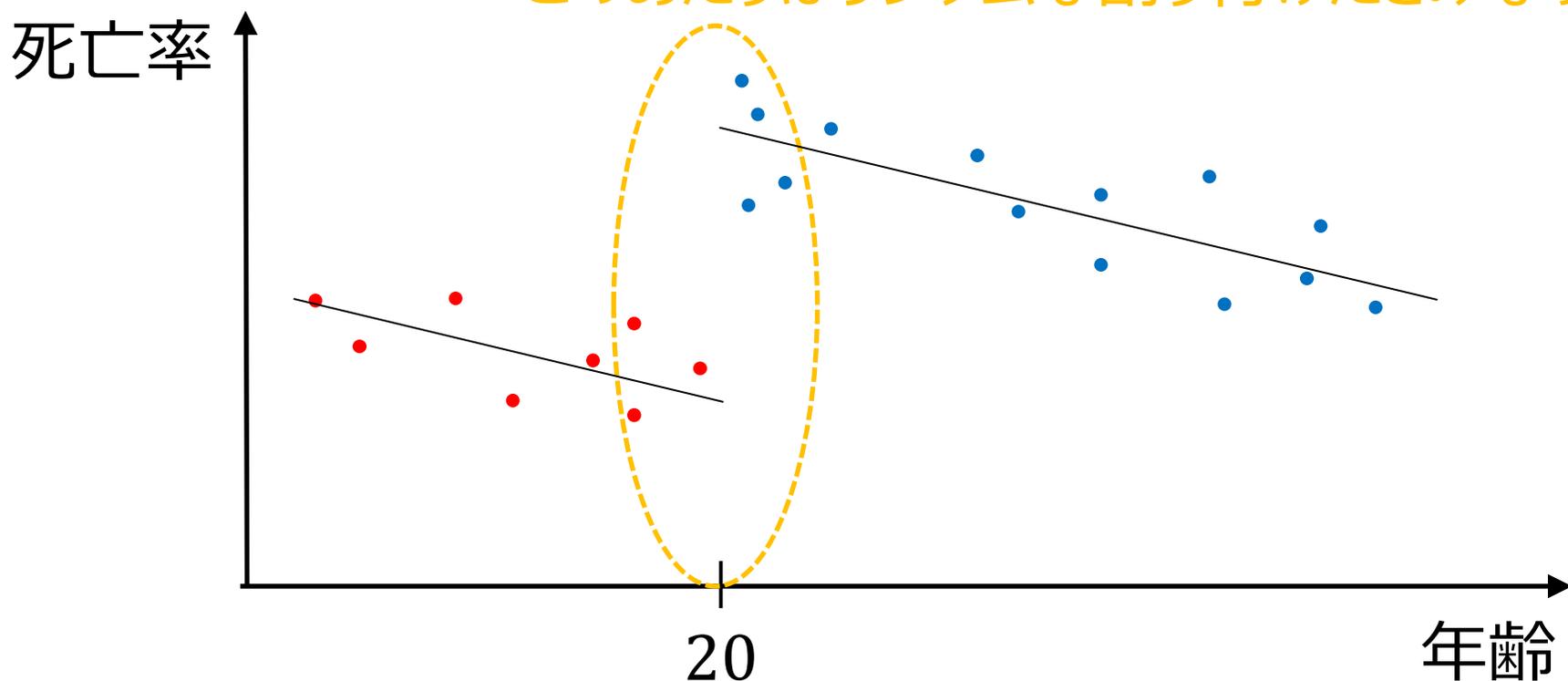
- 例：飲酒は体に悪いか？（飲酒で死亡率が増加する？）
 - ランダムに飲酒グループと禁酒グループを割り付けられない...
- 年齢別死亡率データから因果関係を導けないか？
- あたかもランダムに割り付けられた状況を見つける
- 日本では20歳以上が飲酒可能なので、20歳以上が飲酒グループ（介入群）、20歳未満が禁酒グループ（対照群）
 - これでは恣意的な割り付けになるので、一見ダメそう
 - 19歳から20歳になる瞬間で介入と対照が一気に切り替わる
→ この周辺の人にはランダム割り付けとみなしてよいのでは？

回帰不連続デザインの考え方：

割り付けの境界付近データではRCTが行われたとみなす

- 境界付近では、割り付けがランダムに行われていると仮定する
 - 割り付け以外の変数の分布が同じ

このあたりはランダムな割り付けだとみなす



回帰モデルによる回帰不連続デザイン： 割り付けを表すダミー変数を導入

- 回帰不連続デザインを回帰モデルで実現する
- 割り付けを表すダミー変数（割り付けの閾値 τ ）を導入：

$$D = \begin{cases} 1 & (x \geq \tau) \\ 0 & (x < \tau) \end{cases}$$

- 回帰モデル：

$$Y = \beta X + \alpha + \gamma D$$

- γ ：介入効果をあらわす（介入によるジャンプ）

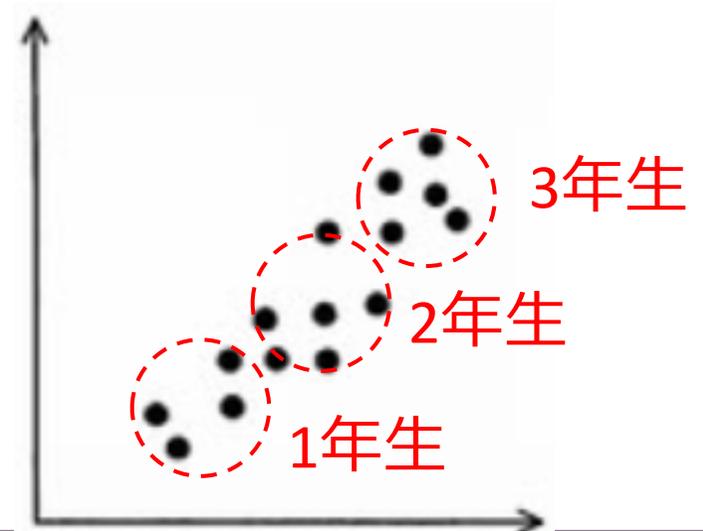
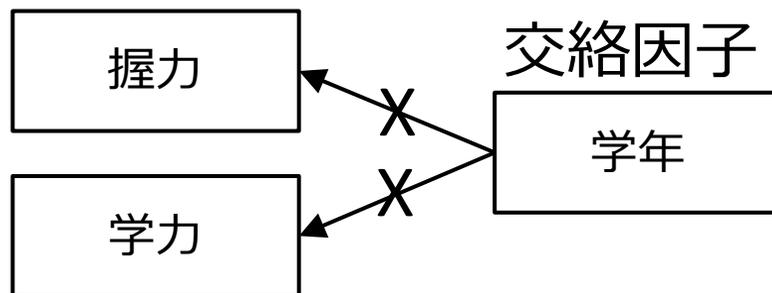
回帰不連続デザインの限界： いくつかの仮定が必要

■ 回帰不連続デザインの仮定：

1. 割り付けルールが適用されない場合、ジャンプはないこと
 - ◆ 回帰モデルの場合、割り付けられなければ、20歳以降も $Y = \beta X + \alpha$ が成立するはず
 - ◆ 確認できる場合とできない場合がある
2. 他の変数に同様のジャンプがないこと
 - ◆ 他の変数が原因になっている？
– タバコとの区別がつかない
3. 割り付けを自分でコントロールできない

層別分析： 交絡因子を固定して効果をみる

- 結果変数の変化は、原因変数の変化によるものか、それとも交絡因子の変化によるものか？
- 握力と学力のケースでは、学年が交絡因子だった
 - 交絡因子を固定すれば、握力と学力は独立になるはず
- 層別解析：交絡因子を固定することによって、交絡因子の影響を除く
 - 学年ごとにデータを分けて相関をみる

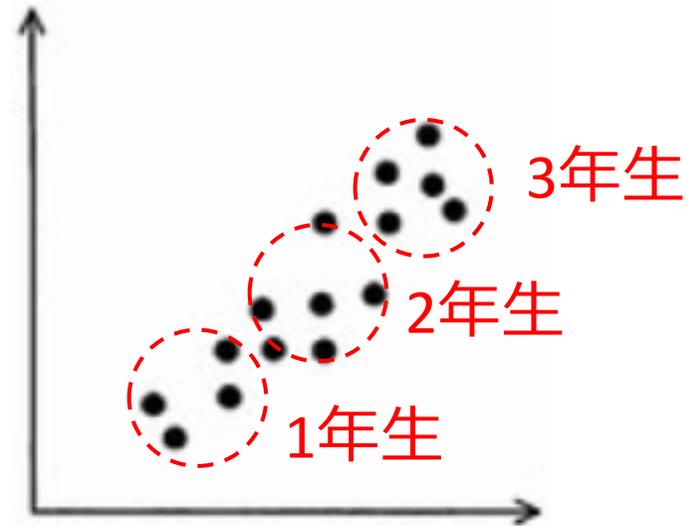


回帰モデルの利用： 交絡因子を説明変数に組み込む

- 回帰モデルの中に交絡因子を取り込む：

$$\text{テストの点} = \beta_1 \times \text{握力} + \beta_2 \times \text{学年}$$

- β_1 が交絡因子（学年）の効果を取り除いた握力の効果になる
 - β_2 は交絡因子（学年）の効果을 吸収する
- 前提：
 - 交絡因子の「あたり」がついている
 - 線形回帰モデルが成り立っている



マッチング

反事実：

反事実がわかれば因果効果が測れる（が不可能）

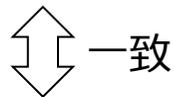
- 全く同じ状況における
 - 介入を受けた場合の結果
 - 介入を受けなかった場合の結果の差を介入の因果効果として定義した
- 反実仮想：「実際には起こらなかったが、もし起こっていたら...」
 - 実際には薬を飲んで風邪が治ったが、もし飲まなかったとしたら...
 - ◆ 治った⇒薬の効果は関係ない（因果関係なし）
 - ◆ 治らなかった⇒薬が効いた（因果関係あり）
- 反事実は観測されない：「あちら側」の世界の出来事



マッチング： 反事実の仮想的実現

- データの中から反事実ペアをみつける
 - (20代, 男性, 京都在住, 介入あり) と
(20代, 男性, 京都在住, 介入なし)は互いに反事実とみなせる
 - これらの結果の差は因果効果とみなしてよいのでは…？
- マッチング：あるデータに対して原因変数以外の（交絡因子を含む）変数がすべて一致しているほかのデータを見つけペアにする
 - 実現されない反事実を仮想的に実現する

(20代, 男性, 京都市在住, 介入あり) → 成績向上



(20代, 男性, 京都市在住, 介入なし) → 成績停滞

この差が
因果効果

反事実ペアの発見： 実際には近いペアで代用することも

- 実際には「全変数の値が一致」しているペアを見つけるのは困難
- 代わりに「概ね一致」しているペアを見つける

(20歳, 男性, 京都市左京区在住, **介入あり**)

⇕ 概ね一致

⇕ 概ね一致

(21歳, 男性, 京都市上京区在住, **介入なし**)

- 2つのデータ $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top$, $\mathbf{y} = (y_1, y_2, \dots, y_D)^\top$ の距離

- $D(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + \dots + |x_D - y_D|$ (絶対距離)

- $D(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \dots + (x_D - y_D)^2}$ (ユークリッド距離)

- $D(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$ (マハラノビス距離)

- ◆ 標準偏差で割って正規化するなども

傾向スコア：

層別分析／マッチングの「次元の呪い」問題を解決する

- 次元の呪い：変数が多いと層別分析／マッチングが困難になる
 - 最新医療の導入と死亡率の間の因果関係の例だと、患者の難病度合いの他、年齢や、病院に機器があるか等の変数がありうる
- 傾向スコア $\Pr[Z = T | X]$ の導入：介入群への割付け確率
 - X ：原因変数・結果変数以外の変数（マッチングに使う変数）
 - たとえばロジスティック回帰モデルによる傾向スコアの定義：

$$P(Z = T | X = \mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \mathbf{x})} = \sigma(\boldsymbol{\beta}^T \mathbf{x})$$

$\boldsymbol{\beta}$ ：パラメータ

- 多変数を1次元の変数（傾向スコア）に縮約するイメージ

傾向スコアの使い方①：

マッチングにおける次元の呪いの解消

- 傾向スコア $\Pr[Z = T \mid X]$ （割付け確率）をどう使うか？
- $\Pr[Z = T \mid X] = 1/2$ であるデータは、どちらに割り付けられるかが50:50 – つまりランダム化試験の結果とみなしてよいのでは？
 - ただし、この条件を満たすデータは多くないだろう...
- $\Pr[Z_i = T \mid X_i] \approx \Pr[Z_j = T \mid X_j]$ となる2つのデータ*i*と*j*があったとすると、*X*からはこれらの区別がつかない
- 傾向スコア（1次元）をマッチングに用いる：
 - 傾向スコアが $\Pr[Z_i = T \mid X_i] \approx \Pr[Z_j = T \mid X_j]$ であり、実際には $Z_i = T$ 、 $Z_j = C$ であった2つのデータ*i*と*j*をマッチングする

傾向スコアの使い方①： マッチングの例

- 傾向スコア（1次元）をマッチングに用いる
- 例：介入（最新医療適用）を説明するモデルを推定：
$$\Pr[Z = \text{適用} \mid X = (\text{患者は難病か, 年齢, 機器の有無})]$$
- $\Pr[\text{最新医療を適用}]$ をマッチングの変数として利用する
（一次元なのでマッチングが行いやすい）
- （ほぼ）同じ割り付け確率をもちながら、実際に介入された人とされなかった人を比較する
- ◆ $\Pr[Z_i = \text{適用} \mid X_i] \approx \Pr[Z_j = \text{適用} \mid X_j]$ であり、実際には $Z_i = \text{適用}$ 、 $Z_j = \text{不適用}$ である2つのデータ i と j をマッチング

傾向スコアの使い方②： 逆確率重みづけ

- 全てのデータが同じ確率 $\Pr[Z = T | X] = 1/2$ で割りつけられたとみなせれば、RCTと同じとみなせるだろう
- いま、 $\Pr[Z_i = T | X_i] = 1/2$ 、 $\Pr[Z_j = T | X_j] = 1/4$ なるデータ i と j があるとすると、 j は i よりも2倍くらい $Z = T$ になりにくい
- これを補正するためには「 j は i の2個分」と考えればよいのでは？
- 逆確率重みづけ：傾向スコアの逆数で重みづける
 - $Z_i = T$ である i は $\frac{1/2}{\Pr[Z_i=T|X_i]}$ 個分と数える
 - $Z_i = C$ である i は $\frac{1/2}{1-\Pr[Z_i=T|X_i]}$ 個分と数える

差の差 (DID)

パネルデータ： 介入前後でデータがある場合の因果推論

- パネルデータ = クロスセクションデータ + 時系列データ
 - クロスセクションデータ：ある時点において、複数の項目を集めたデータ
 - 時系列データ：ひとつの項目を複数時点集めたデータ

	2014	2015	2016	2017
売り上げ	100	200	300	200
気温	29℃	33℃	35℃	31℃

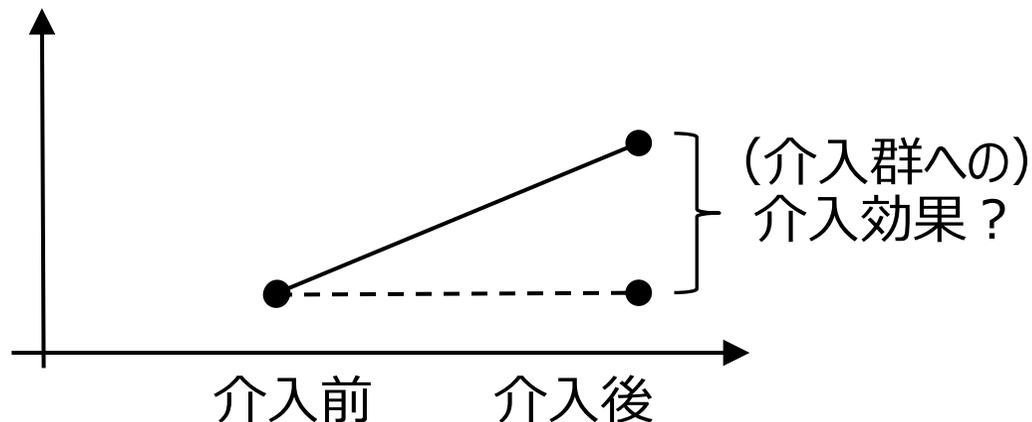
時系列データ

クロスセクションデータ

- 因果推論の文脈では介入前後のデータがある場合を指す

パネルデータからの因果推論： 介入前後のデータ差から効果を測るのは難しい

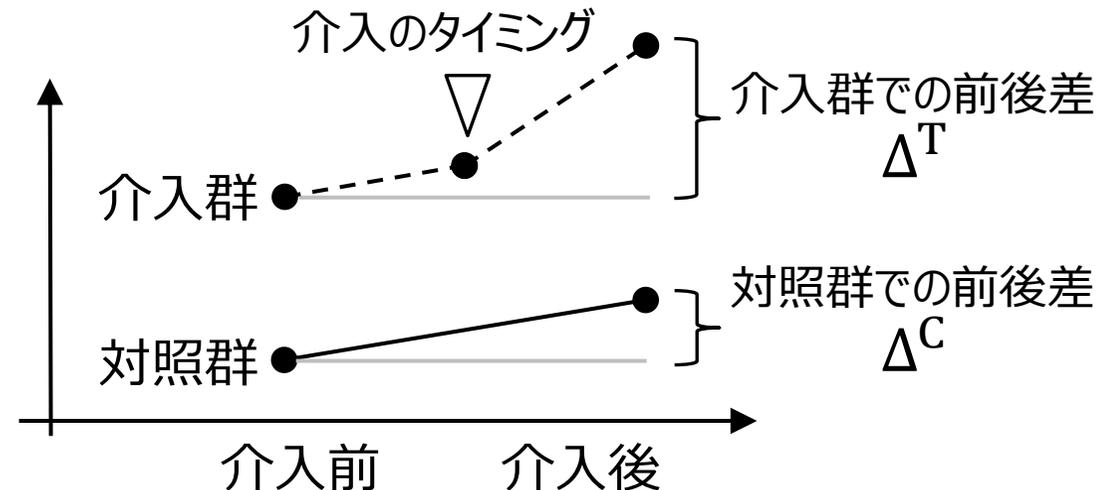
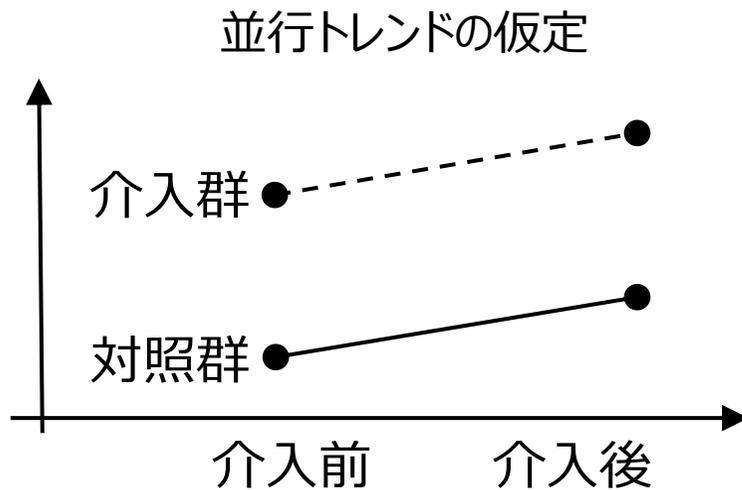
- あるグループに対して、介入前後のデータが得られたとする
- 介入前後の差は介入効果といってもよいか？
 - 「介入しなかった場合には変化がない」という仮定の下ではこれは正しい
 - そうでない場合には成り立たない



差の差 (DID) 法 :

並行トレンドを仮定し介入前後のデータから効果を測る

- 並行トレンドの仮定 : 介入が無かったとしたら、介入群と対照群は同じように変化する
- 介入効果があるなら、介入タイミング後に結果が改善する
- 介入効果 : $\Delta^T - \Delta^C$ (介入群・対照群それぞれの前後差の差)



差の差 (DID) 法： 回帰モデルによる実現

- 回帰モデル： $Y = \alpha D + \beta T + \gamma DT$
 - $D \in \{0,1\}$ ：介入の有無を表す変数
 - $T \in \{0,1\}$ ：介入タイミングの前後を表す変数
 - γ ：介入効果（介入タイミングより後における介入群にのみ効果）
- 仮定：
 - 並行トレンド仮定：介入が無ければ両群は同様に変化する
 - 共通ショック仮定：
 - ◆ 介入前後に、結果に影響を与える他のイベントが起こっていない
 - ◆ 仮に起こったとしても、その効果は両群において同じである

今回の話題： 因果推論

- 相関関係と因果関係は異なるという話：相関 \neq 因果
- ランダム化試験（RCT）：因果関係を導く最も正しいやり方
- 準試験：データから因果を導くにはどうしたらよいか
 - 交絡因子が観測できる場合：層別解析、回帰モデル、マッチング、傾向スコア
 - できない場合：差の差（パネルデータを前提）、回帰不連続デザイン（割り付けが境界線付近でランダムとみなせる必要）

異常検知

障害や災害：

大規模で複雑なシステムの障害は一旦起こると大損失

生産ラインの故障



コンピュータの
ウィルス感染・不正侵入



ほか、カード不正使用、オレオレ詐欺、監視カメラ、…

監視システム： 障害の予兆を早期発見できれば対処可能

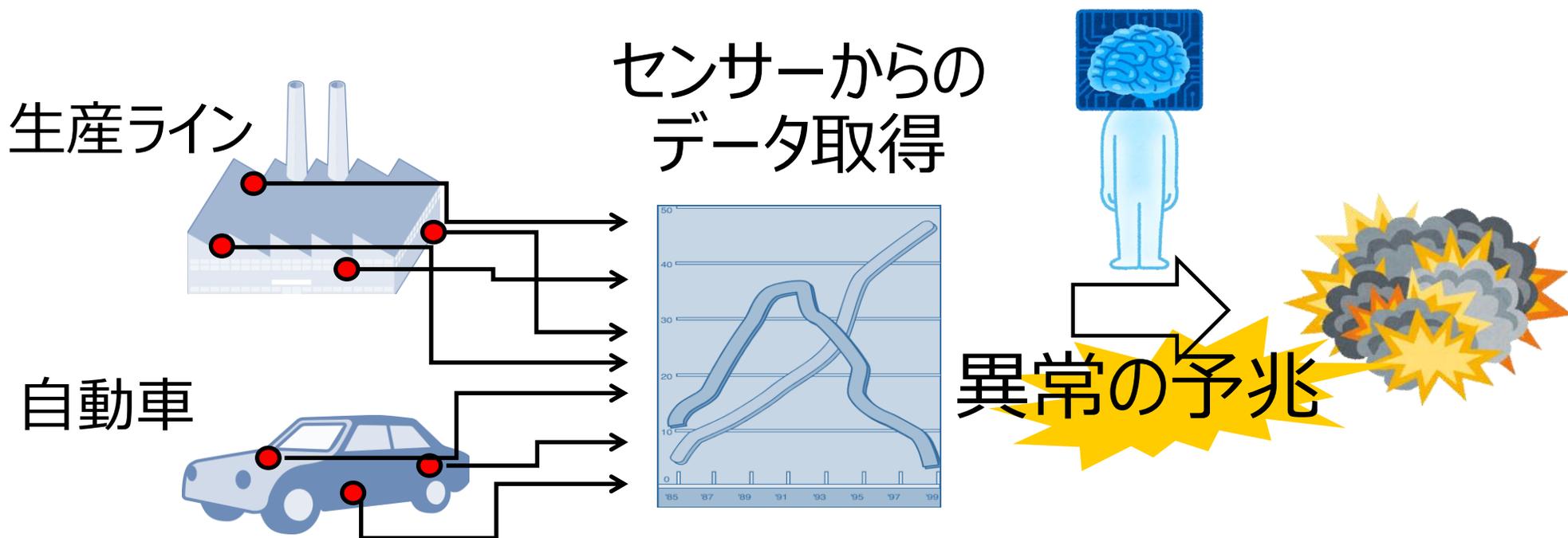
まだ間に合う！



異常検知：

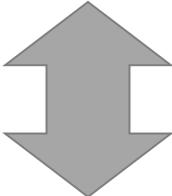
システムの問題を事前に察知し知らせる

- 設置されたセンサーからの取得データをもとに異常を早期検知したい



異常検知の難しさ：

障害のデータがないため、判別モデルが適用できない

- 障害発生時のデータが集められる場合：
 - 予め検出したいタイプの異常がわかっている場合
 - ある程度の頻度で起こる場合→ 判別モデル（ロジスティック回帰等）が適用可能
- 
- しかし、重大な障害ほど初めて出会うものが多い
- 過去のデータがない（判別モデルが使えない）

異常検知の考え方：

正常時を捉え、そこからの逸脱を見つける

■ 文書や会話：

- 正常：よく出てくる単語・表現
- 異常：めったに出てこない単語・表現

■ センサー：

- 正常：普段の値の範囲
- 異常：普段では出てこない値



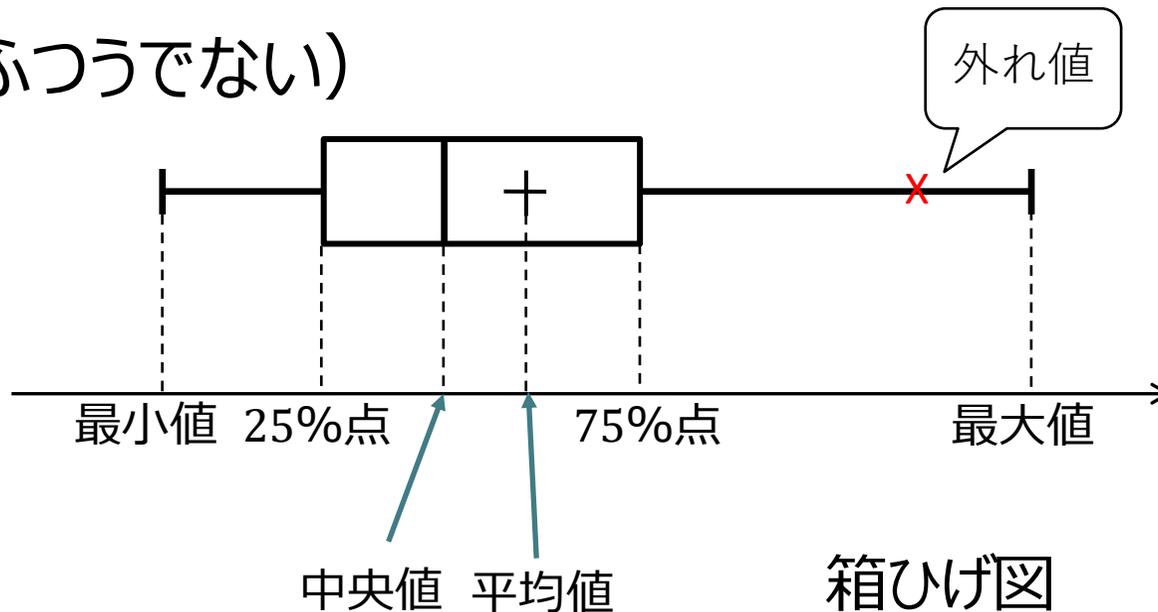
異常検知技術：

「普通でない」データのパターンをみつけたす技術

- 異常：データの中に含まれる「普通でない」パターン
 - 対象のシステムのなんらかの異常を原因として表れる
 - ◆ クレジットカード不正使用・システム侵入・テロ
・システムダウン...
 - あるいは対象システムの状態変化
 - ◆ 新規ニューストピック出現・システムの設定変更
・環境変化...
- これらをデータの中から発見し報告するのが異常検知

異常検知の基本的な考え方： 稀な値を異常と考える

- 単純な場合として1変数（例えば温度）を対象とした異常検知を考える
- 通常の変の範囲をとらえる（温度は通常20～50°Cの間）
- そこから逸脱した値を検出し、異常として報告する
(80°Cはふつうでない)



統計的な異常検知：

モデルからの生成確率が小さいデータを異常と考える

- 正常なデータ x_1, x_2, \dots, x_n が与えられる
- 正常なデータからモデル $p(x)$ を推定する
 - たとえば最尤推定によって（正常時モデルの）パラメータ推定値を得る
- 検証対象のデータ x^{NEW} に対して、モデルが与える確率 $p(x^{\text{NEW}})$ を算出する
- これがある閾値 τ より低ければ $(p(x^{\text{NEW}}) < \tau)$ 異常として報告する

多次元データのモデル： 多次元正規分布の最尤推定

- データが多次元の場合は、たとえば多次元正規分布：

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

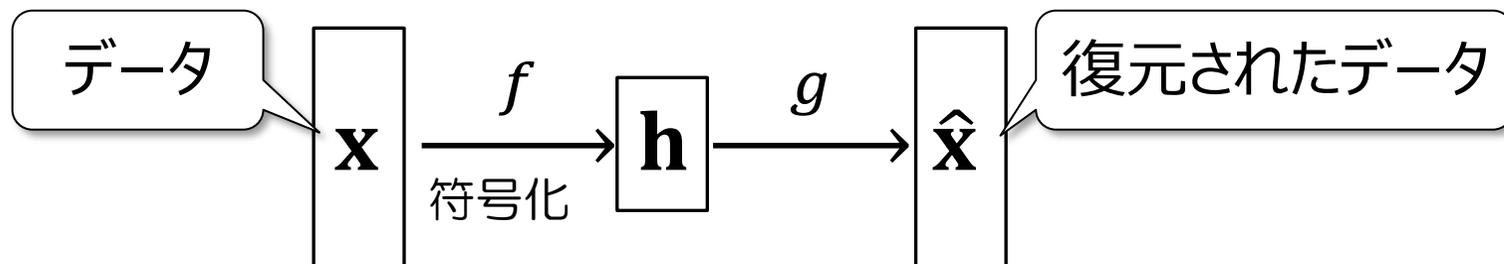
- 正常データ $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$ からパラメータを最尤推定

- 平均： $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$

- 分散共分散行列： $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}}) (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$

再構成誤差を用いた異常検知の考え方： モデルによる再構成誤差が大きいデータを異常と考える

- 符号化 $\mathbf{h} = f(\mathbf{x})$ と復号化 $\hat{\mathbf{x}} = g(\mathbf{h})$



- 符号化器はデータを低次元に射影する
- 復号化器はもとの空間に戻す
- データ \mathbf{x} が正常なデータなら小さな誤差で復号できるはず
 - 例：ユークリッド距離の場合： $e(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - g(f(\mathbf{x}))\|_2^2$

再構成誤差を用いた異常検知の考え方： 再構成誤差は本質的なデータ表現に載るノイズ

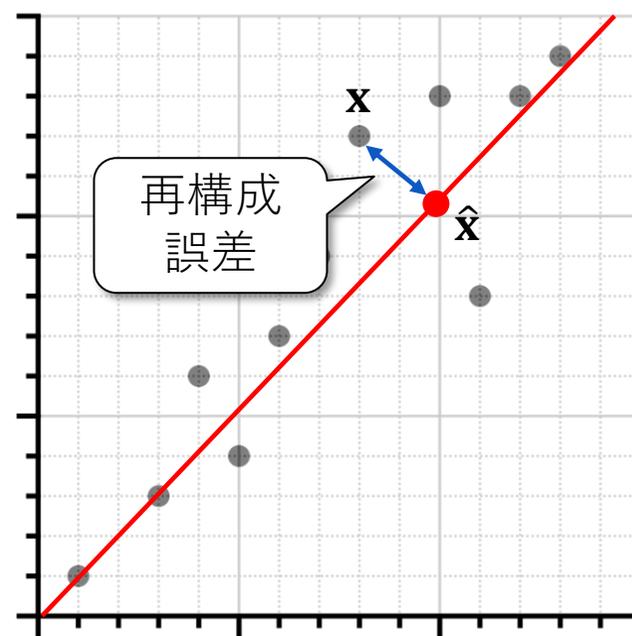
- 低次元空間がデータの本質的な形を捉えているとすると、再構成誤差は「ノイズ」として解釈できる
 - 大きいノイズを含むものを外れ値とする
 - 射影の距離が再構成誤差にあたる

- 再構成誤差 $\| \mathbf{x} - \hat{\mathbf{x}} \|_2^2$ が正規分布：

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2} \| \mathbf{x} - \hat{\mathbf{x}} \|_2^2\right)$$

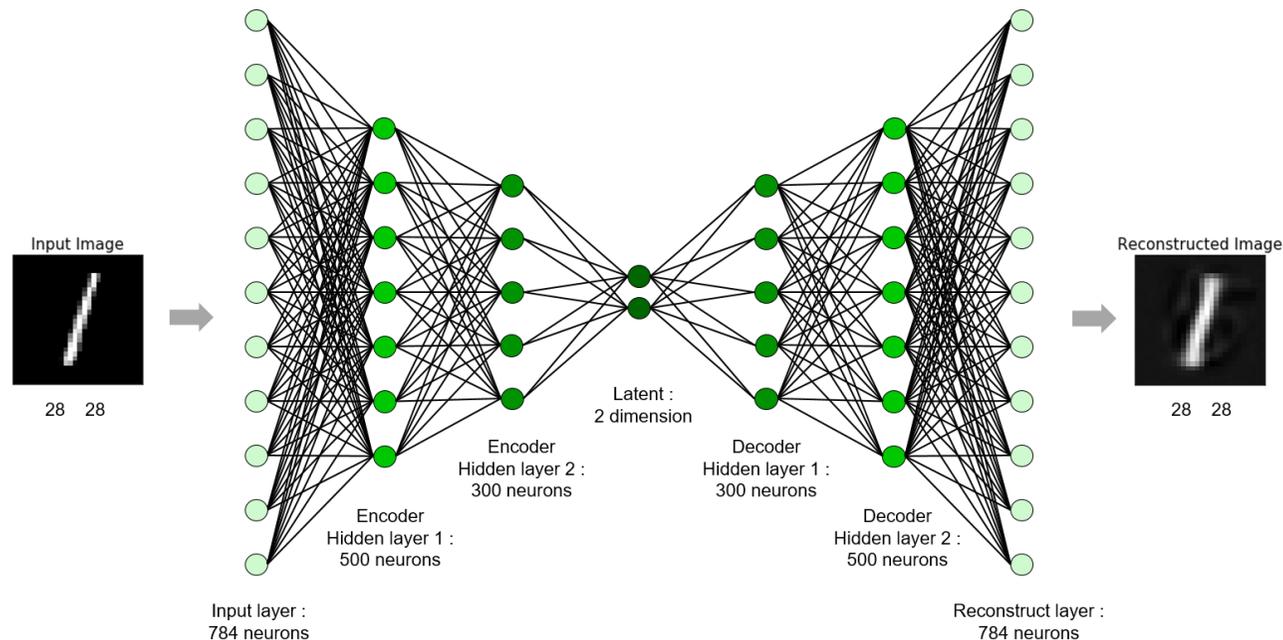
に従うとして $p(\mathbf{x}^{\text{NEW}}) < \tau$ で判定すると
前述の枠組みで解釈できる

2次元データを1次元に射影した例



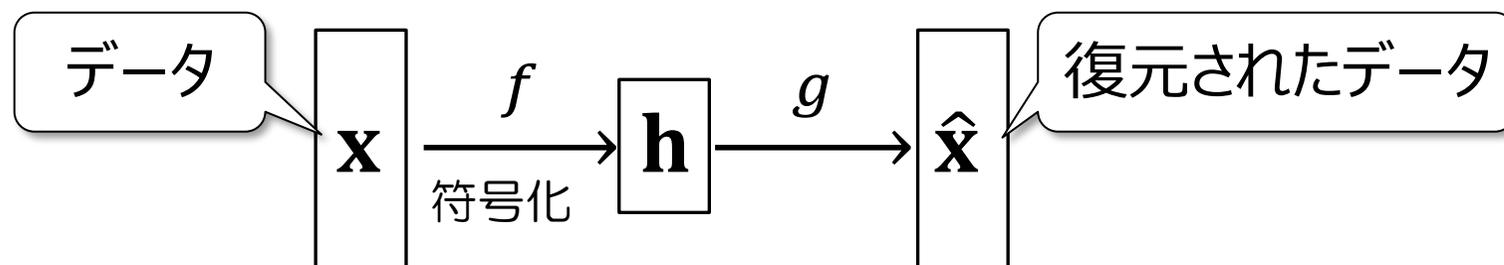
ニューラルネットによる符号化・復号化： 自己符号化器（オートエンコーダ）

- 符号化器 $\mathbf{h} = f(\mathbf{x})$ と復号化器 $\hat{\mathbf{x}} = g(\mathbf{h})$ をどう実現するか
- 主成分分析（PCA）：符号化 $\mathbf{h} = \mathbf{F}\mathbf{x}$ ・ 復号化 $\hat{\mathbf{x}} = \mathbf{G}\mathbf{h}$ とした場合（ \mathbf{F} と \mathbf{G} は行列、つまり線形変換）
- 自己符号化器は符号化・復号化をニューラルネットでおこなうもの



自己符号化器による異常検知： 自己符号化器の再構成誤差を異常度とする

- 符号化 $\mathbf{h} = f(\mathbf{x})$ と復号化 $\hat{\mathbf{x}} = g(\mathbf{h})$



- 正常なデータから符号化器・復号化器を推定：

$$(f, g) = \operatorname{argmin}_{f, g} \sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$$

- 新たなデータ \mathbf{x}^{NEW} の再構成誤差を異常度とする：

$$\|\mathbf{x}^{\text{NEW}} - \hat{\mathbf{x}}^{\text{NEW}}\|_2^2$$

まとめ： 異常検知

- 異常検知：システムの障害は一旦起こると大損失であり早期検出したい
- 異常データは稀なので、正常時をモデル化する
 - 多次元正規分布、ニューラルネットワーク（オートエンコーダ）
- データの「珍しさ」によって異常度を判定
 - 確率モデルからの生起確率の小ささによってデータの異常度を判定
 - 再構成誤差の大きさによってデータの異常度を判定

まとめ

統計的モデリングの考え方： データから母集団を知る

- 部分（データ）から全体（母集団＝データの生成機構）を知る
- モデルのさまざまな使い方
 - 予測（将来のデータについて）
 - モデルそのものから知識を得る
 - ◆ 回帰モデルでどの独立変数が使われるか
 - ◆ 因果分析
 - データの検証
 - ◆ 異常発見
 - データの生成：近年盛ん

そのほかのトピック：

本講義で紹介できなかった（が重要な）トピック

- 潜在変数モデル：決して観測されない変数
 - トピックモデル、主成分分析・因子分析
 - EMアルゴリズムによる推定
- 独立でない構造をもったデータのモデリング：時系列、ネットワーク構造
- 計算統計（おもにベイズモデリングの実施のための技）
 - マルコフ連鎖モンテカルロ・変分推論
- 応用：推薦システム、機械学習

アンケートの実施：

Kulasisからアンケートを実施してください

- アンケートを実施してください：
統計的モデリング基礎（鹿島久嗣）を選択
- スマートフォン、ノート又はタブレット型パソコンで実施
 - 持っていない方は、授業後、サテライト演習室や自宅のパソコンからアンケートに答えてください
- このアンケートは、授業について受講者の皆さんからの意見を聴き、授業・教育環境の改善に役立てようとするためのものです。
- アンケートは無記名方式であり、回答内容が成績評価に影響することはありません