2022.4.27
スタートアップゼミ #5

アプローチの基礎

因果推論・多段階最適化

M1 増橋 佳菜

Set-up：因果推論用語と統計学

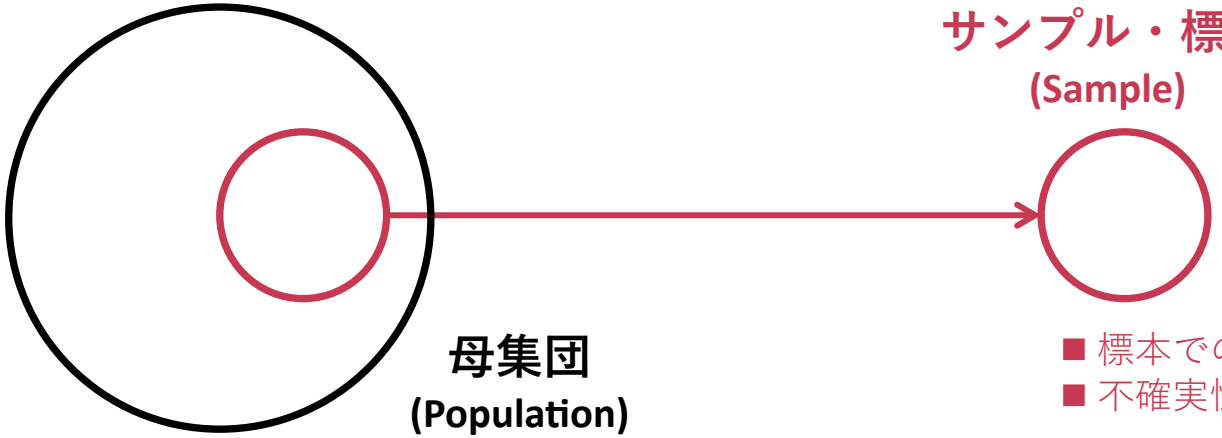
アウトカム：結果変数 Y（例：血圧/経路選択結果）
(Outcome = Endpoint = Dependent Variable)

曝露因子：処置の有無(割当て)を表すの2値変数 $X \in \{0,1\}$（例：薬の投薬、カーナビ情報の提供）
(Exposure = Treatment = Explanatory/Independent Variable)
※ **処置 = 介入**

共変量：交絡因子（←経済学）、調整変数とも。全部まとめてUまたは U_1, U_2, U_3
(Covariates = Adjustment Variable)

推定と推論

母集団における関心のある値 (parameter) を知りたい



- 標本での計算（**推定; estimation**）
- 不確実性の考慮（**推論; inference**）

期待値: $E[Y]$（母集団全体におけるYの期待値）

条件付き期待値: $E[Y|U]$（Uが同じ値の集団におけるYの期待値）

$$X = \begin{cases} 1 & \text{処置（有）} \\ 0 & \text{処置（無）} \end{cases}$$

セレクションバイアス

セレクションバイアス とは：

原因として考えている変数(処置変数)とアウトカムの関係が、想定する因果関係以外に存在する状況

1. サンプルセレクションバイアス

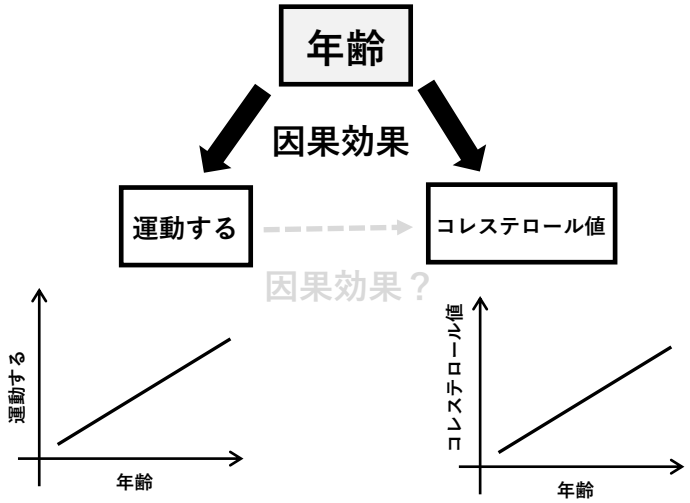
母集団からかけ離れた代表性のないサンプルをとる場合に生じる

→ 対策：「全て」から**無作為**にサンプルをとる

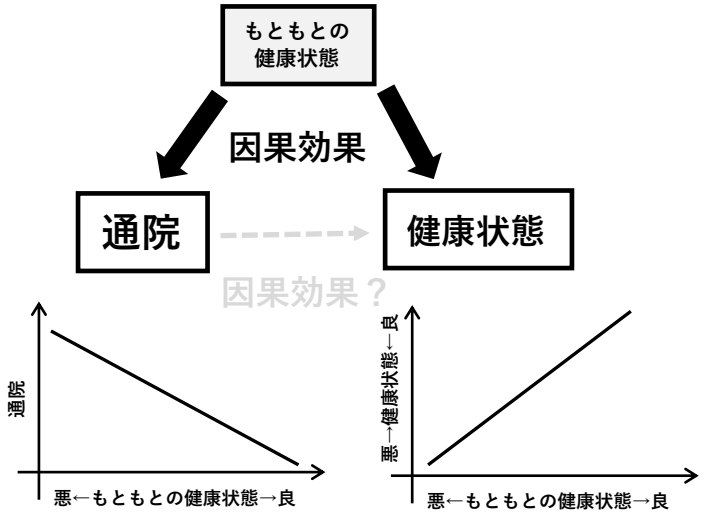
2. セルフセレクションバイアス

個人が自らの意思によって行動を選択した結果、ある行動を取る人たちのグループと取らない人たちのグループの間で特性の差が生じる

例1：運動する程コレステロール値が上がる???



例2：病院に行く人ほど不健康になる???



Rubinの潜在的結果アプローチ

因果推論の根本問題 (Fundamental Problem of Causal Inference) (Paul; 1986)

同一主体は処置の有無に関していずれか一方の結果しか観測できないため、
実現しなかった潜在的結果（**反事実**）は非観測となり、そのままでは有無比較が不可能。

個人レベルで考える

現実の世界

薬を飲んでいる

ID 服薬あり $X = 1$
血圧 $Y = 120$

薬を飲まなかった

ID 服薬なし $X = 0$
血圧 $Y = 140$

反事実の世界

2 種類の処置

ID 服薬あり $X = 1$
血圧： $Y^{x=1}$

ID 服薬なし $X = 0$
血圧： $Y^{x=0}$

服薬Xが血圧Yに与える（この個人における）

因果効果 $= Y^{x=1} - Y^{x=0}$
$= 120 - 140 = -20$

反事実
→ **観測不能**

集団レベルで考える

現実の世界

薬を飲んだ人 $(X = 1)$と
飲まなかった人 $(X = 0)$

反事実の世界

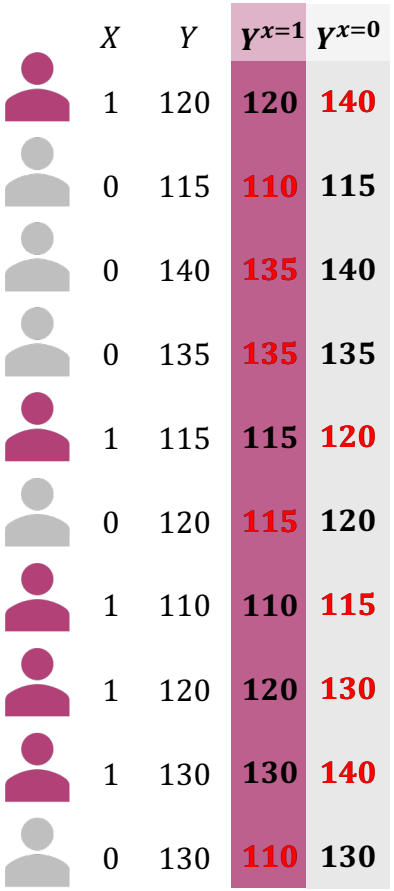
全員が$X = 1$

全員が$X = 0$

因果効果とは

平均因果効果 (Average Treatment Effect; ATE)

服薬の有無Xが血圧Yに与える（集団レベルでの）平均としての因果効果 $E\left[Y^{x=1}\right] - E\left[Y^{x=0}\right]$



X	Y	$Y^{x=1}$	$Y^{x=0}$
1	120	120	140
0	115	110	115
0	140	135	140
0	135	135	135
1	115	115	120
0	120	115	120
1	110	110	115
1	120	120	130
1	130	130	140
0	130	110	130

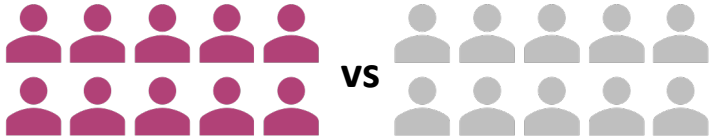
$$ATE = E\left[Y^{x=1}\right] - E\left[Y^{x=0}\right] = -8.5$$

推定値

120　128.5

限界効果 Marginal Effect

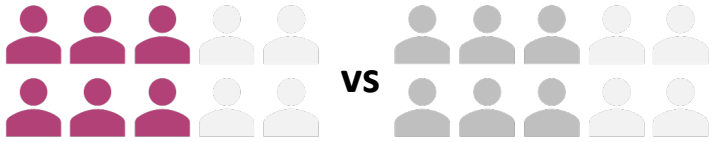
$$E\left[Y^{x=1}\right] - E\left[Y^{x=0}\right]$$



母集団の全員が薬を飲んだ$(X = 1)$場合と全員が飲まなかった場合$(X = 0)$の比較

条件付き効果 Conditional Effect

$$E\left[Y^{x=1}|U\right] - E\left[Y^{x=0}|U\right]$$



母集団のうち**”U層の中で”**（例:女性のみ）全員が薬を飲んだ$(X = 1)$場合と全員が飲まなかった場合$(X = 0)$の比較

Marginal Effect と Conditional Effect の関係

Marginal Effect = **Conditional Effect** の重み付け平均

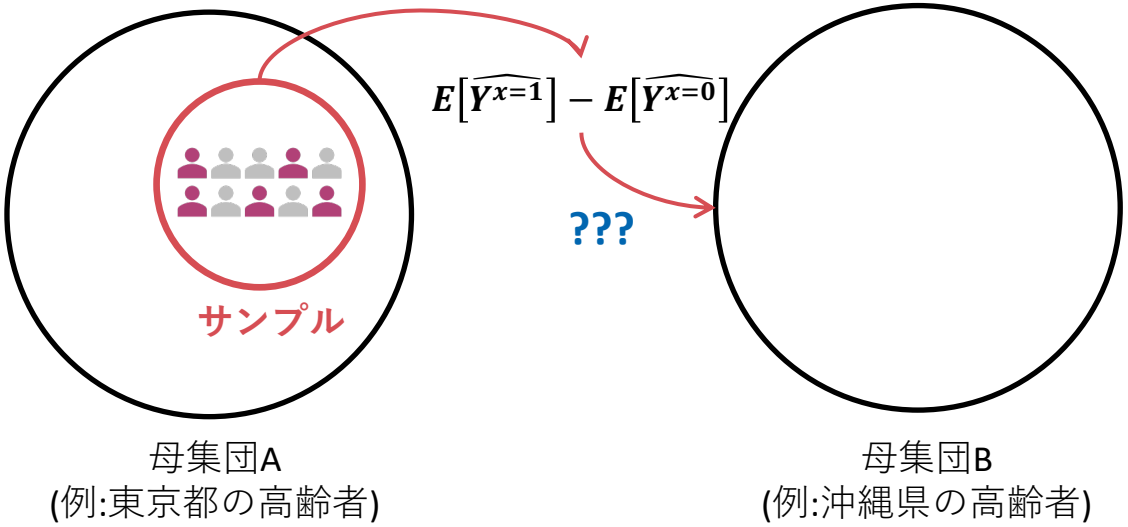
共変量が 2 値変数の場合 ※説明はこっちで

$$E[Y^{x=1}] - E[Y^{x=0}] = \{E[Y^{x=1}|X=0] - E[Y^{x=0}|X=0]\} * Pr(X=0)$$
$$+\{E[Y^{x=1}|X=1] - E[Y^{x=0}|X=1]\} * Pr(X=1)$$

一般化表記

$$E[Y^{x}] - E[Y^{x^*}] = \sum_{x} \{E[Y^{x}|X=x] - E[Y^{x^*}|X=x]\} * Pr(X=x)$$

外的妥当性 :サンプルから推定された効果が他の母集団への一般化が可能か



※ 効果は個人属性に基づいて異なりうる（サンプルの**異質性**の考慮）

AssociationとCausation (数理的理解)

データ

Causation

知りたいもの

$$E\left[Y^{x=1}\right] - E\left[Y^{x=0}\right]$$

Association

これで代用可能?

$$E[Y|X=1] - E[Y|X=0]$$

これが知りたい！

$$E[Y^x] = \{E[Y^x|X=1] * Pr(X=1) + E[Y^x|X=0]\} * Pr(X=0)$$

薬を飲んでいるひとの平均 / 薬を飲んでいるひとの割合 / 薬を飲んでいないひとの平均 / 薬を飲んでいないひとの割合

もし、 $E[Y^x|X=1] = E[Y^x|X=0]$ なら、

$$E[Y^{x=1}] = E[Y^{x=1}|X=1](= E[Y^{x=1}|x=0])$$

$$E[Y^{x=0}] = E[Y^{x=0}|X=0](= E[Y^{x=0}|X=1]) \quad \cdots(1)$$

$$\because E[Y^{x=1}] = \{E[Y^{x=1}|X=1] * Pr(X=1) + E[Y^{x=1}|X=0]\} * Pr(X=0)$$

もし、 $E[Y^x|X=x] = E[Y|X=x]$ なら(1)から、

$$E[Y^{x=1}] = E[Y|X=1]$$

$$E[Y^{x=0}] = E[Y|X=0]$$

因果効果の識別 (Identification)

もし、 $E[Y^x|X=1]=E[Y^x|X=0]$ なら

もし、 $E[Y^x|X=x]=E[Y|X=x]$ なら

$$E[Y^{x=1}]=E[Y|X=1]$$

$$E[Y^{x=0}]=E[Y|X=0]$$

Causation **Association**

$$\boldsymbol{E[Y^{x=1}]-E[Y^{x=0}]=E[Y|X=1]-E[Y|X=0]}$$

反事実の世界
（データから観測不能）

現実の世界
（データから観測可能）

仮定1： $E[Y^x|X=1]=E[Y^x|X=0]$ (**Exchageability ; 交換性**)

仮定2： $E[Y^x|X=x]=E[Y|X=x]$ (**Consistency ; 一致性**)

仮定3： **Positivity ; 正値性 = 処置の有無の確率が共に0ではない**

仮定1 : Exchageability ; 交換性

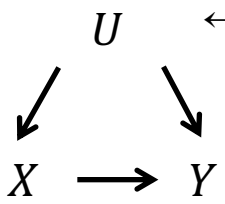
アプローチ

1. **無作為化比較試験（RCT試験）**
2. **自然実験**
3. **調整**

・非ランダムな処置の割り当てを許容する
・ Exchageabilityが成立しない原因を特定
・医学研究・社会科学でよく用いられる

Exchageabilityが成立しない主な原因

1. 交絡

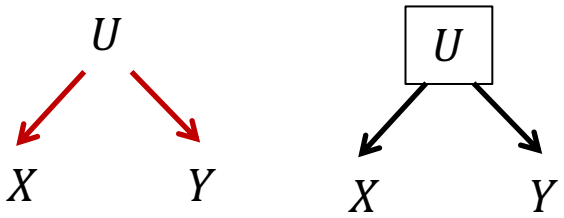


← **交絡変数 U ：**
従属変数 Y と独立変数 X 両方に相関する外部変数

2. 選択バイアス(=セレクションバイアス)

⇨ **DAG (Directed Acyclic Graph ; 有向非巡回グラフ)**
を用いて要因間の関係性を考える

DAGルール 1 : “共通の原因”による裏口経路(バックドア)



・UはXとYの**共通の原因(Common Cause)**

・UとYの間には統計的な関連が生じる

・ $Y \leftarrow U \rightarrow X$ という**裏口経路(Backdoor Path)**

・ Uの値を揃える(=条件付け)ことで裏口経路を閉じる

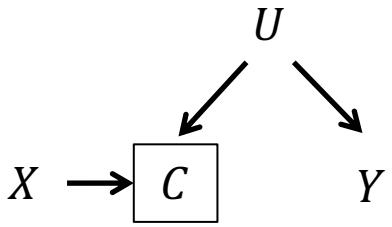
○ 交絡 (Confounding)

共通原因により生じる因果効果に由来しない関連

○ 交絡因子 (Confounder)

条件付けによって裏口経路を閉じることができる要因

DAGルール 2 : “共通の効果”による裏口経路(バックドア)



・CはXとの**共通の効果(Common Effect ; Collider)**

・ Cを条件付けると$X-$Uの間に関連が生じる

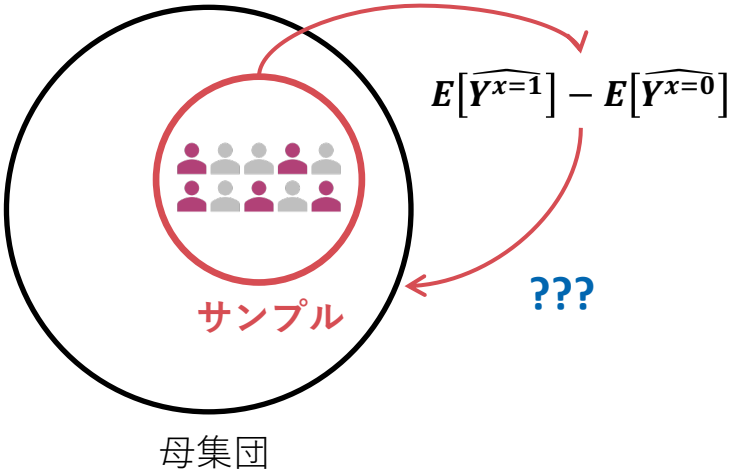
・ Cを条件付ける＝Cの値が同じサンプル(人)を対象に分析する

・ **Collider Stratification Bias ; 選択バイアス**

・選択バイアスは**内部妥当性**の問題

内部妥当性：

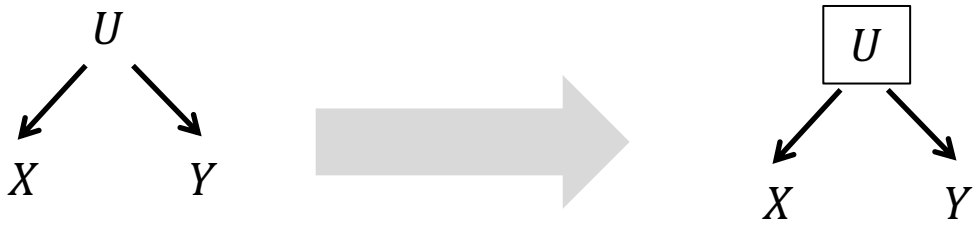
観察データから推定された効果が真の因果効果を捉えられているか



選択バイアスが生じるシナリオ

1. 追跡の失敗 (Loss-to-follow-up)
2. 競合リスク (Competing Risk)
3. サンプリング方法
4. 欠測データ
5. 自己選択

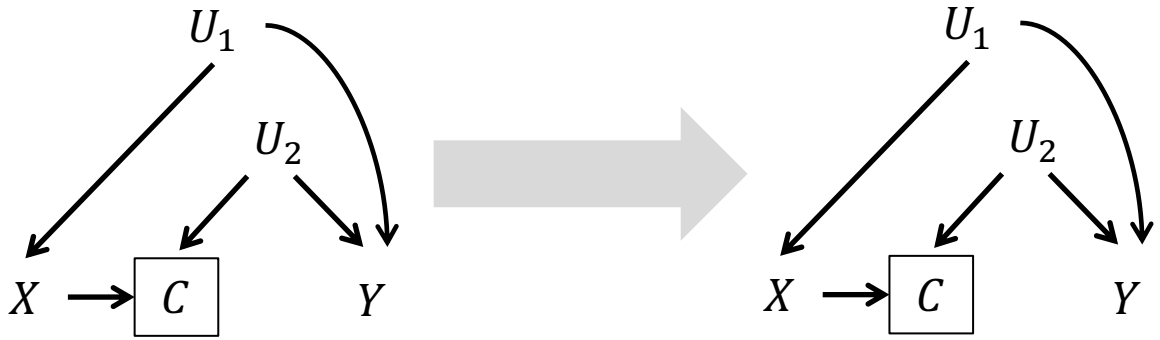
DAGルール 3 : 全ての経路を閉じる



条件付け(層化)

Uの値が同じサンプルを分析

例：



・２つの裏口経路
・ U_1による交絡
・Cによる選択バイアス

・ U_1 U_2を条件付け
・裏口経路なし

Exchangeabiluty
Conditional on $U_1$and U_2

Conditional Exchangeability: 条件付き交換性

- $U(U_1, U_2, U_3, \dots)$を条件付けると裏口経路なし
- **Conditional Exchangeability** : $E[Y^x|X=1,U] = E[Y^x|X=0,U]$

（Uの層内で）反事実世界のアウトカム=観測されたアウトカム

$$\boldsymbol{E[Y^{x=1}|U] - E[Y^{x=0}|U] = E[Y|X=1,U] - E[Y|X=0,U]}$$

p.4 参照

$$\because E[Y^x] = \{E[Y^x|X=1,U] * Pr(X=1|U) + E[Y^x|X=0,U]\} * Pr(X=0|U)$$
$$= \{E[Y^x|X=x,U] * \{Pr(X=1|U) + Pr(X=0|U)\}\}$$
$$(\because \boldsymbol{[Y^x|X=1,U] = E[Y^x|X=0,U]})$$
$$= E[Y^x|X=x,U]$$
$$= E[Y|X=x,U](\because \boldsymbol{Consistency})$$

Marginal	Conditional
Exchangeability : $E[Y^x\|X=1] = E[Y^x\|X=0]$	**Exchangeability** : $E[Y^x\|X=1,U] = E[Y^x\|X=0,U]$
推定できる因果効果：	(条件付きで)推定できる因果効果：
$\boldsymbol{E[Y^{x=1}] - E[Y^{x=0}] = E[Y\|X=1] - E[Y\|X=0]}$	$\boldsymbol{E[Y^{x=1}\|U] - E[Y^{x=0}\|U] = E[Y\|X=1,U] - E[Y\|X=0,U]}$ **層化**

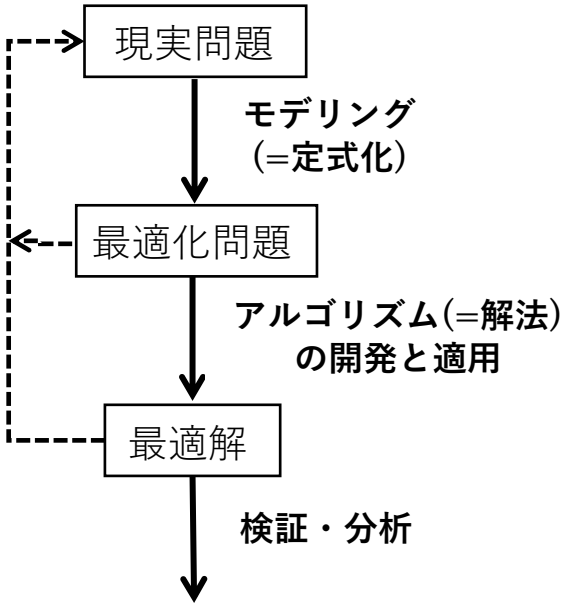
交通分野の最適化

- **制約条件**を満たす解の中で**目的関数**を最小(最大)にする解を求める問題＝**最適化問題**
- 最適化は都市のデザイン(＝設計)を決定するための一つの手段として多く用いられる

$$\min_{x} f(x) \quad subject\ to \quad g(x) = 0, h(x) > 0$$

制約条件

- **制御変数** x
- **目的関数** f

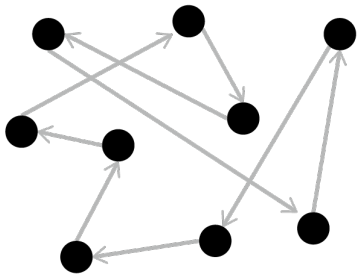


[参考] 交通分野における最適化の例

	制御変数	目的関数	制約条件
配送計画	配送経路	旅行時間の最小化	配送場所
バス路線計画	路線選定	収益の最大化 乗客の効用の最大化	車両台数
シェアモビリティシステムの設計	ポート位置	収益の最大化	モビリティの配置数
公共空間の設計計画	施設配置の位置	受益者の利益の最大化 まちなかの回遊性の最大化	利用可能な敷地 予算制約

NP困難な問題

巡回セールスマン問題（Traveling Salesman Problem）



与えられるデータ　n個の地点と各2地点間の距離
条件　すべての地点をちょうど１回ずつ経由して元の地点に戻る（＝巡回路）
目標　巡回路の**総距離を最小**にする

巡回路の選び方は $n! = n(n-1)(n-2)\ldots 2 \cdot 1$ 通り

ナップザック問題（Knapsack problem）

与えられるデータ　n個の品物(選択肢)の利用価値(効用)と重量及び上限重量
条件　選んだ品物の重量の合計がナップザックの重量上限を超えない
目標　品物(選択肢)の**利用価値の合計(=効用の和)を最大**にする

各商品に対して選ぶか否かの2通りの選択があるので、組み合わせ数は 2^n通り

これらは、**全組み合わせを列挙**すれば原理的には解ける。

しかし…

指数関数的にnの増加に伴って選択肢が増加→組み合わせ爆発が起こる **NP困難**

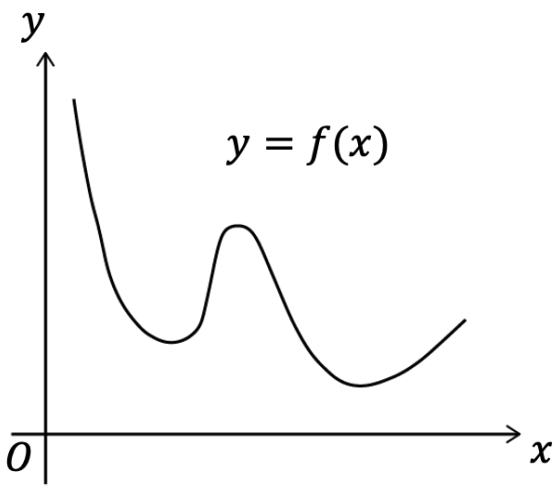
参考：組み合わせ爆発 https://youtu.be/Q4gTV4r0zRs

組合せ(離散)最適化問題

最適化問題の中でも最適解の集合が**離散的**（あるいは離散的なものに減らせる）な問題

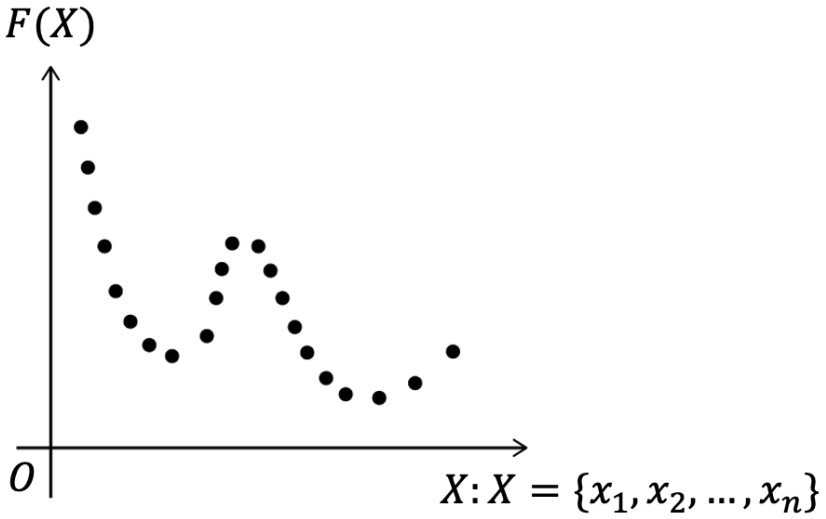
連続的

微分による関数の最小化が可能



離散的

個々の離散変数の組み合わせによって解が変化
→ 最適解も離散的に変化



基本的な概念

- 線形計画問題(LP)は多項式時間で解ける
- 0/1線形計画問題(0/1-LP)や整数計画問題(IP)は指数関数時間→ **NP困難**

0/1-LPやIPの問題の定式化を、LPで解けるように**条件を緩和（線形緩和）**

線形計画法への緩和

線形計画法(LP)の標準形　LPにはいくつかの形があるが、全てこの標準形で表現可能

(等式)標準形

実数上の任意の入力

$$A = [a_{ji}]_{j=1,\dots,m,i=1,\dots,n}$$
$$b = (b_1, \dots, b_m), c = (c_1, \dots, c_n)$$

に対して、制約条件

$$A \cdot X = b$$
$$\leftrightarrow \sum_{i=1}^{n} a_{ji} x_i = b_j \ (j = 1, \dots, m)$$
$$\text{かつ} \ \ x_i \geq 0 (i = 1, \dots, n)$$

のもとで

$$X \cdot c^T = \sum_{i=1}^{n} c_i x_i$$

を**最小化**する問題

(不等式)標準形

入力 A, b, c に対して、制約条件

$$A \cdot X \leq b$$
$$\leftrightarrow \sum_{i=1}^{n} a_{ji} x_i \leq b_j \ (j = 1, \dots, m)$$
$$\text{かつ} \ \ x_i \geq 0 (i = 1, \dots, n)$$

のもとで

新しい変数(余裕変数, Slack Variable) y_j を導入

$\sum_{i=1}^{n} a_{ji} x_i + y_j = b_j$ かつ $y_j \geq 0$

$$B = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \vdots & a_{2n} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & 0 & 0 & \cdots & 1 \end{pmatrix}$$
$$d = (c_1, \dots, c_n, 0, \dots, 0)$$

入力 B, b, d の**(等式)標準形**になる

線形計画法への緩和

例：**ナップザック問題（Knapsack problem）**の線形緩和

データ　n個の品物の利用価値(効用)はc_iと重量はω_i
条件　ナップザックの重量上限はb
目標　品物の**利用価値の合計(=効用の和)を最大**にする

$\omega_1, \omega_2, \dots, \omega_n, c_1, c_2, \dots, c_n, b$

n個のブール変数: $x_1, x_2, \dots, x_n$
※ 0 or 1 の2値変数

$$x_i = \begin{cases} 1\ (i\text{番目の品物を詰める}) \\ 0\ (i\text{番目の品物を詰めない}) \end{cases}$$

このとき、解きたい問題は、

$$\sum_{i=1}^{n} \omega_i x_i \leq b \quad \text{かつ} \quad x_i \in \{0,1\} \ (i = 1, \dots, n)$$

の 制約条件のもとで、

$x_i \in \{0,1\}$を$x_i \geq 0$に緩和すれば
LPの不等式標準形が得られる

$$\sum_{i=1}^{n} c_i x_i \quad \text{を最大化すること}$$

すなわち、

$$\sum_{i=1}^{n} (-c_i) x_i \quad \text{を}\mathbf{最小化}\text{すること}$$

NP困難 → 線形計画問題(LP)は多項式時間で解ける

動的計画法 Dynamic Programming : DP

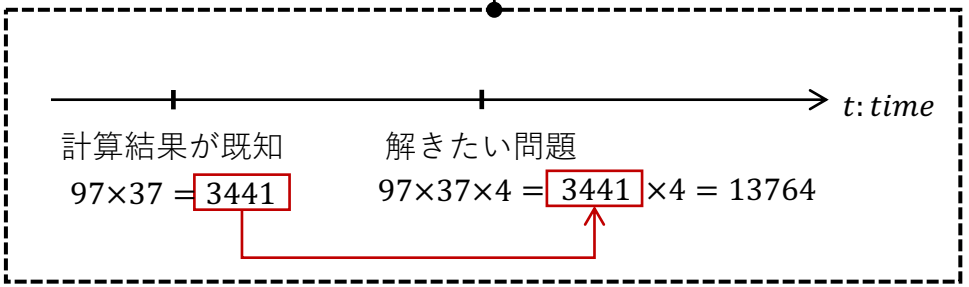
動的計画法 (Bellman, R.E.(1953))

解きたい問題をいくつかの簡単な問題に分割して、**多段階最適化問題（＝動学的最適化）**として解を得るアルゴリズムの総称。

※ ここでは**離散**最適化を扱う

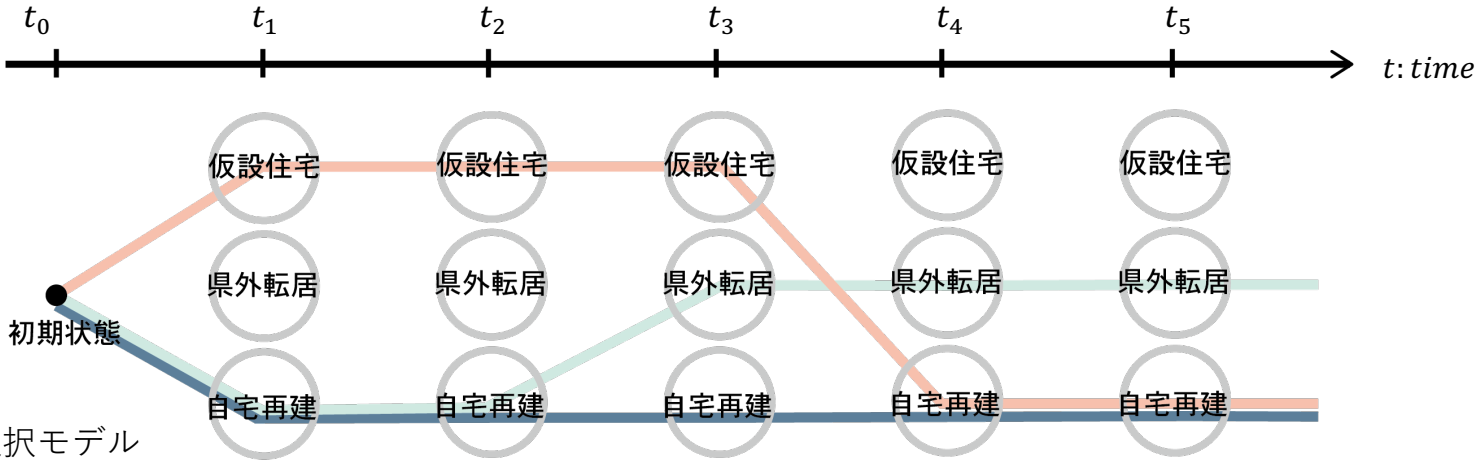
→ **動的離散選択モデル** を推定

現在の状態から**未来**を算出可能



再帰的計算過程が必要

- 将来状態の計算
- 状態遷移を確率的に与えることも可能
- 終端時刻Tが決まっている必要

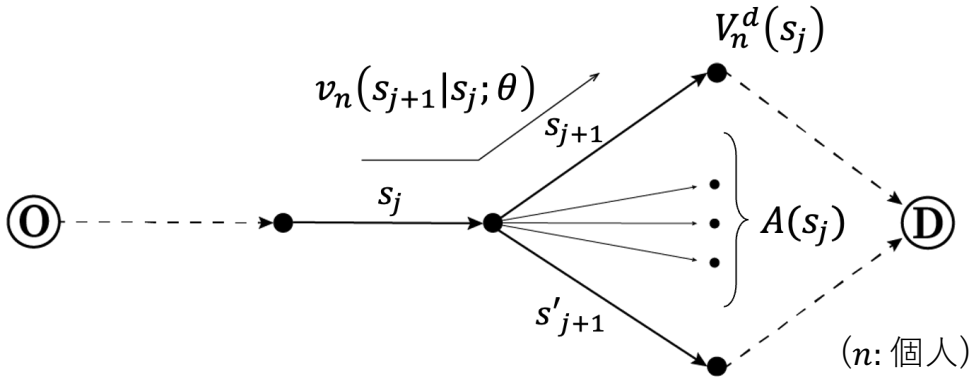


例：居住選択モデル

動的計画法 Dynamic Programming : DP

再帰的計算過程が必要

- 将来状態の計算
- 状態遷移を確率的に与えることも可能
- 終端時刻Tが決まっている必要



Discounted Recursive Logit Model
(Oyama & Hato ; 2017)

状態価値関数（Bellman方程式）

$$V^d(s_j) = E\left[\max_{s_{j+1}\in A(s_j)} \{v(s_{j+1}|s_j;\theta) + \mu\epsilon(s_{j+1}) + \beta\, V^d(s_{j+1})\}\right]$$

今期の効用　　次期の期待効用

- 将来の価値関数は時間割引を考慮した効用の最大化
- 今期と次期の状態により価値関数を定義
- 次々期以降の効用は次期の期待価値関数の中に含んでいる

β: 時間割引率 ($\beta \in (0,1)$)

$$u(s_{j+1}|s_j;\theta) = v(s_{j+1}|s_j;\theta) + \mu\epsilon(s_{j+1})$$

今期の効用

選択確率

$$P(s_{j+1}|s_j;\theta) = \frac{\exp[v(s_{j+1}|s_j;\theta) + \beta\, V^d(s_{j+1})]}{\sum_{s_{j+1}'\in A(s_j)} \exp[v(s_{j+1}'|s_j;\theta) + \beta\, V^d(s_{j+1}')]}$$

→ **最尤法** で推定

s_t: 時刻tの状態
V: 価値関数(目的関数)
d: 吸収状態
μ: スケールパラメータ
θ: モデルパラメータ

二段階最適化問題

ある問題の最適化が別の問題の最適化の**制約条件**になっている問題

上位問題

$$\max_{\mu} V(x, \mu)$$
$$subject\ to\ G(x, \mu) \leq 0, H(x, \mu) = 0$$

μ: 上位側の制御変数
(計画者の政策変数)

↑ **制約条件**

下位問題

$$\max_{x} f(x, \mu)$$
$$subject\ to\ g(x, \mu) \leq 0, h(x, \mu) = 0$$

x: 下位側の制御変数
(利用者の行動結果(効用最大化))

例：近藤卒論 2021

上位問題

避難滞在場所 場所選択モデル

場所の避難者の収容人数、移動距離、親戚知人宅

↑ **制約条件**
選択肢集合を限定

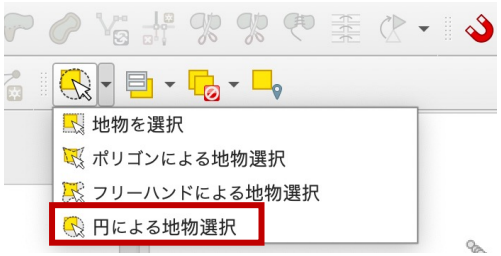
下位問題

避難滞在場所 選択肢集合形成モデル

標高、川からの距離、累積雨量

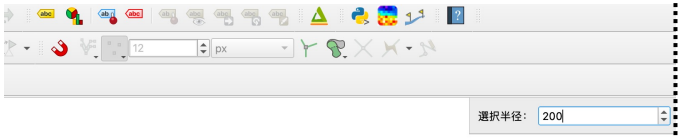
Tips 1 : QGISで地物を範囲抽出する

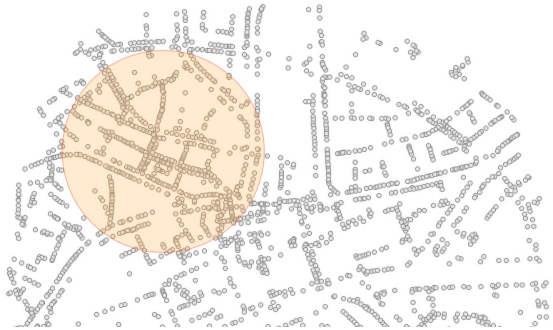
1. さまざまな地物の選択方法がある



円による地物選択は
抽出中心と半径を任意に設定できて便利

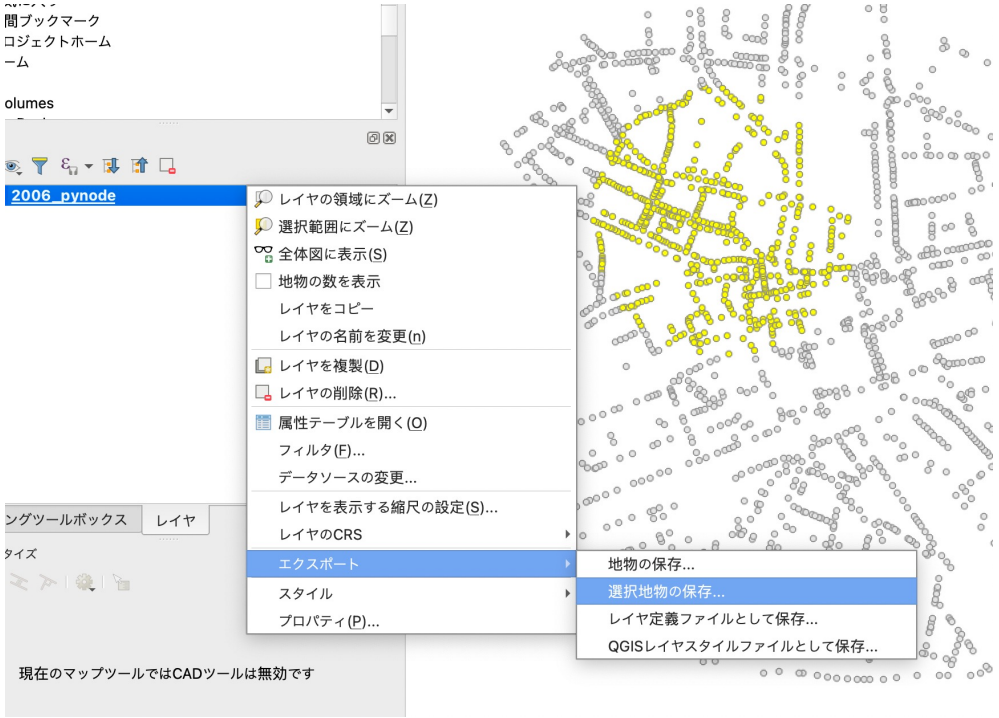
2. 抽出半径を指定して抽出できる





抽出の中心にしたい部分をクリックしたら、
右上で半径を設定してEnter

3. 選択した地物だけのファイルを保存しておく



こんなとき

・特定の半径の地物だけ抽出したいとき

・**PPデータを毎回全部importすると<u>QGISの動作が重くなる</u>とき**

Tips 2 : pandas DataFrame　　for index,row in df.iterrows():

```
path_df = path_df.reset_index()
link_df = link_df.reset_index()

pre_TripID = link_df.iloc[0, 1]
# pre_TripIDの初期値は(Path/)Link データの一行目のTripID
print('preTripID:',pre_TripID)

L_TripID = []   # TripIDの最終リスト
L_k = []   # 注目リンクkの最終リスト
L_a = []   # 遷移先リンクaの最終リスト
segmentTripID = []   # TripIDのリスト
segmentL_k = []   # TripIDごとの注目リンクkのリスト
segmentL_a = []   # TripIDごとの遷移先リンクaのリスト

for index, row in link_df.iterrows():
    if row.TripID == pre_TripID:
        # pre_TripIDに等しい間はTripIDごとのリストを作成
        id = row.TripID
        k = row.Link
        a = row.Link
        segmentTripID.append(id)
        segmentL_k.append(k)
        segmentL_a.append(a)
    else:   # TripIDが次のものに変わった瞬間に
        segmentTripID.pop()   # 末尾を削除
        segmentL_k.pop()   # 末尾を削除
        segmentL_a.pop(0)   # 先頭要素を削除

        # 最終リストにTripID=pre_TripIDのsegmentを追加
        L_TripID.extend(segmentTripID)
        L_k.extend(segmentL_k)
        L_a.extend(segmentL_a)
```

indexについてのエラーが出た場合，
DataFrameのindexの振り直しで解決することが多い

for index,row in df.iterrows():

indexにはその行のindex番号、rowにその行の要素が返る
このとき row.列名 または row[‘列名’]で各列の要素を取り出せる

［注意点］複雑な構造を処理の背後に含むため**遅い**!!

こんなとき

・使い道はたくさん！
・入れ子構造の場合に書きやすい
・csvを1行ずつ処理したい場合

参考文献

多段階最適化

- 夏合宿ゼミ2015 資料 http://bin.t.u-tokyo.ac.jp/summercamp2015/document/key_urata.pdf
- 理論輪読会2014 資料 http://bin.t.u-tokyo.ac.jp/rindoku14/#1-3
- 論文ゼミ2013 http://bin.t.u-tokyo.ac.jp/rzemi13/test/今泉2.pdf
- 組合せ最適化入門：線形計画から整数計画まで https://www.slideshare.net/shunjiumetani/ss-17197023?from_action=save
- 簡単そうで難しい組合せ離散最適化 https://www-or.amp.i.kyoto-u.ac.jp/files/open-campus-04.pdf
- Rust, J., Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher, Econometrica, Vol.55, pp.999-1033, 1987.
- 大山雄己, 羽藤英二（2017）一般化RLモデルを用いた災害時の経路選択行動分析, 交通工学論文集, 第3巻, 第 5 号, 1-10

因果推論

・UNBOUNDEDLYの記事 https://www.krsk-phs.com/entry/causalinference_lecture_notes

・Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, Vol. 81, No. 396, pp. 945–960, 1986.

・セレクションバイアスとRubinの因果モデル(理論)
http://www.ner.takushoku-u.ac.jp/masano/class_material/waseda/keiryo/R14_sbias.html