

第3章 Network Sampling and Model Fitting

Ove Frank

担当: 森

Outline

- アプリケーション
- 標本調査
- ネットワーク調査の変数
- ネットワーク調査の標本選択
- 確率的ネットワークモデル
- 2項依存性を持ったネットワークモデル


3.1 Applications

標本調査としてのネットワーク

- 薬物中毒者ネットワーク
 - Heroin users, multidrug users, drug injectors...etc
 - Abusers's recruiting routes, needle sharing....
- 性感染症・HIVネットワーク
- 犯罪ネットワーク
- Social capital

3.1-3.2 標本調査

標本調査法: ある集団についての特徴を知るのに母集団に属する一部の対象のみを調べ、関心のある特徴を推計する方法
例) 官庁統計調査, 世論調査, 市場調査...etc

- 標本(sample): 抽出された対象の全体
- 母数(parameter): 推計すべき母集団の特徴
- 推定量(estimator): 標本から構成される母数に対する推計量
- 不偏推定量(unbiased estimator): 推定量のうちその期待値が母数に一致するもの。(一般に誤差を含む, 標本設計に依存)
- 十分統計量(sufficient statistic): 母数の推計に必要な統計量
例) 母平均 μ , 母分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従う母集団から n 標本をとり, μ を標本平均で推定するとき, 標本平均は母数 μ (= μ) に関する十分統計量となる
- 標本抽出にあたっては標本設計を行う  モデルベースアプローチ
 - 無作為抽出, 復元抽出, 層化抽出, etc

3.1-3.2 標本調査(つづき)

- 最小十分統計量
 - ラベルと値: $t = \{(j, y_{-j})\}_{j \in S}$
 - 順序と重複情報はなくともよい
- 不偏推定量の例 (母数平均)
 - Rao-Blackwell $e1(t) = \sum_j y_{-j}/m$
 - Horvitz-Thompson $e2(t) = \sum_j (y_{-j}/N_{-j})$
(N_{-j} は要素が抽出される確率: 包含確率)
- 最小十分統計量は不完全なので最適な推定量を見つけるのは難しい 値に対する何らかのモデルが必要

3.3 ネットワーク調査の変数

標本設計による調査: 要素間の関係を考慮していない
c.f. 地理的に近い要素は似たような特徴を持っている
友達同士は共通した値を持っている
このような要素間の関係を表す変数を用いる
e.g. 人々のコンタクトの頻度, 都市間の物流量
このときに要素はグラフのノード, 要素ペアはエッジ(orアーク)
要素の変数=ノード変数 x , 要素ペアの変数=エッジ y (アーク z)変数
ノード i, j ($x_i, x_j, y_{ij}, z_{ij}, z_{ji}$)

自由度 $d = a^2bc^2$ or $abc(ac+1)/2$

3.4 ネットワーク調査の標本選択

エッジやアークのサンプリング
共犯データ, メールや電話の送受信者, ...etc

観測データ:
 $\{(i, x_i): i \in s\}$ and $\{(i, j, y_{ij}, z_{ij}): i \in s, j \in s\}$

不偏推定量の例 (Horvitz-Thompsonによるアークの値の和)
 $\sum_{i \in s} \sum_{j \in s} \frac{y_{ij} z_{ij}}{s_i s_j}$

Snowball調査の場合, サンプル s_1 におけるノード i の
包含確率 $s_i = P(i \in s_1)$ は s_0 が少なくとも一つのノードを
 $B(i)$ と共通にもつ確率であらわされる

しかしSnowballのような標本設計の包含確率を決定するのは難しい
モデルアプローチへ

3.5 確率的ネットワークモデル

- ノードラベルへの依存性
標本設計に基づく推定量に単一の最適性がない
母集団のモデルを導入
- ランダムネットワークモデル
 - Uniform model
 - Bernoulli model
 - (G, p) : 確率 p でエッジを選択
 - $(G, \{p_{ij}\})$: 確率 p_{ij} でエッジを除去, p_{ij} で置き換え
- 確率ネットワークモデル
 - Holland-Leinhardt model
 - p^* -model
 - マルコフグラフ model

3.6 2項依存性を持ったネットワークモデル

- ノード変数 x_i の確率分布
- x_i の値の条件確率
- ノード間の依存性(条件確率)
- エッジとアークのdyad変数 y_{ij}, z_{ij}, z_{ji} の確率分布
- ノード変数のdyad変数の依存性(条件確率分布 x_i, x_j に依存)
- 図参照