サンプリング

京都大学大学院医学研究科
社会健康医学系専攻
医療統計学分野

大森　崇

目標

- サンプリングの考え方と種類
- 疫学研究とサンプリング
- サンプリング法が使用されている例

Q. サプリメントの購入

- 京都府の高校に通う高校生の
  サプリメントの購入実態を把握したい
  - 週にいくらサプリメントに使うかを調査

どのように調査する?

サプリメントの購入調査

- 京都府の高校の基礎データ(H.19年度)

学校区分	学校数	生徒数
国立	1	602
公立	64	42,184
私立	41	28,650
計	106	71,436

ターゲット集団（target population）

- 研究の目的としている集団
  - 正確にこの集団からの情報を得たい
  - 例）京都府の高校生（71,436人）

Q. ターゲット集団全員を調査

- 全数調査の利点と欠点は?
  - ターゲット集団について、
    知りたいことを完全に把握できる
  - 時間とコストがかかる

集団の一部を調査

- サンプリング（sampling）
  - 明確にわかっているターゲット集団の一部を調べる
  - ある程度の精度で全体像を把握可能
  - ただし偏りを除く工夫が必要

Q. 偏りが入るおそれは

- 近所の高校での調査では？
- インターネットを通じた調査では？

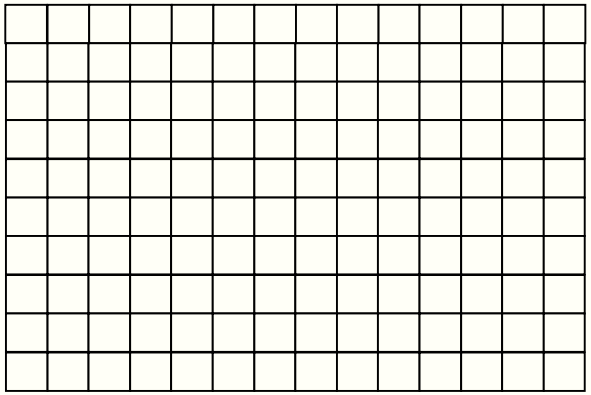
偏りを除く工夫

- ランダム（**randomly**）にサンプリング
  - ➢ ターゲット集団のミニュチュアを作る

もっとも単純なサンプリング法

- 単純なランダムサンプリング
  （Simple random sampling；SRS）
  - わかりやすい方法
  - 他のサンプリング法の基礎

単純ランダムサンプリングのイメージ



ターゲット集団

Q. どうやって選ぶ？

- ランダムにサンプルするために、
  具体的には何が必要？
  - ターゲット集団のリスト
  - 例）京都府の高校生（71,436人）の
    リスト

単純なランダムサンプリング

- サンプルされた集団の測定値からの平均値
  - ターゲット集団の平均値（真値）のよい推定値
  - 推定値はばらつく
  - 標準誤差は推定値の精度の指標

推定の精度（標準誤差）の要因

- サンプリング数
- ターゲット集団でのばらつき
- サンプリング割合（f） sampling fraction
  - ➢ サンプリング数 / ターゲット集団の数

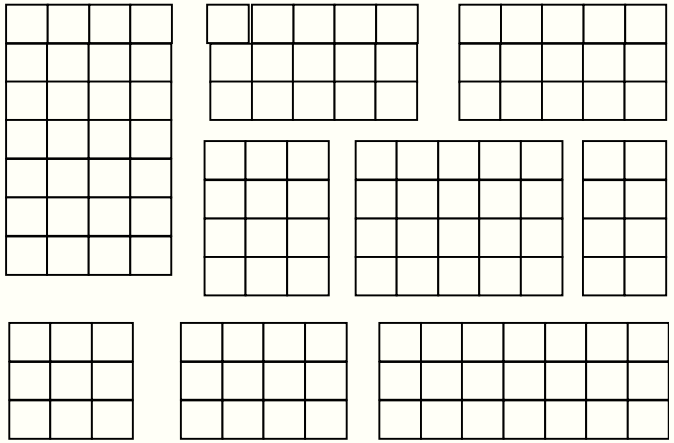
集団に関する情報の利用

- ターゲット集団について
  ある種の情報を利用可能な場合がある
  　例）性別、年齢、地域、なんらかの種類
- このような情報を利用することで
  精度を高めることができることがある

層に分ける（stratification）

- 層化サンプリング stratified sampling
- 調査前のターゲット集団の情報で
  グループ（層）を作成
- 各層からランダムにサンプリング

層化サンプリングのイメージ

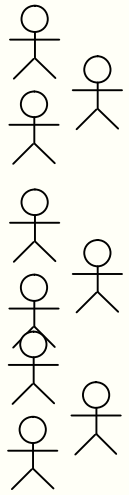


ターゲット集団

サプリメントの購入調査の例

- 学校区分を層とすると

学校区分	生徒数
国立	602
公立	42,184
私立	28,650
計	71,436



それぞれの層からサンプリング

層化サンプリングによる推定

- それぞれの層の重み weight を算出
  - ターゲット集団での各層の人数が必要
- 算出した重みを使った重み付平均

層化サンプリングの利点

- 各層で f が一定のときには
  平均値の計算は非常に簡単
- （fが一定のときには）
  SRSより推定値の精度は悪くならない

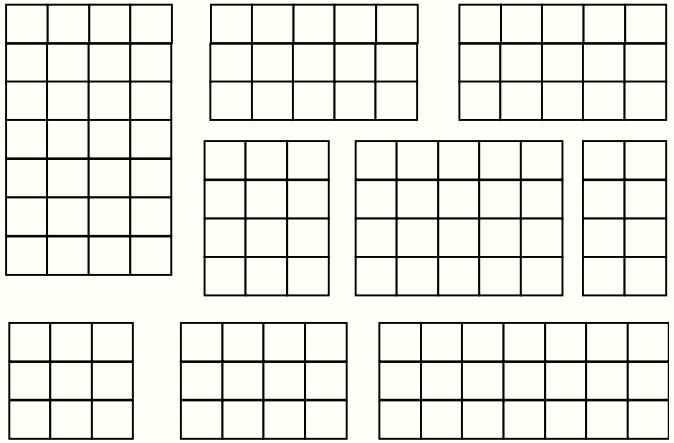
SRSや層化サンプリングの欠点

- ターゲット集団からの
  対象者をランダムにサンプリング
  - ➢ しばしば現実的でない

クラスター cluster を使う

- クラスターサンプリング cluster sampling
- 個人ではなく、グループ（クラスター）をランダムにサンプリング
- クラスターの中の全員を調査

クラスターサンプリングのイメージ

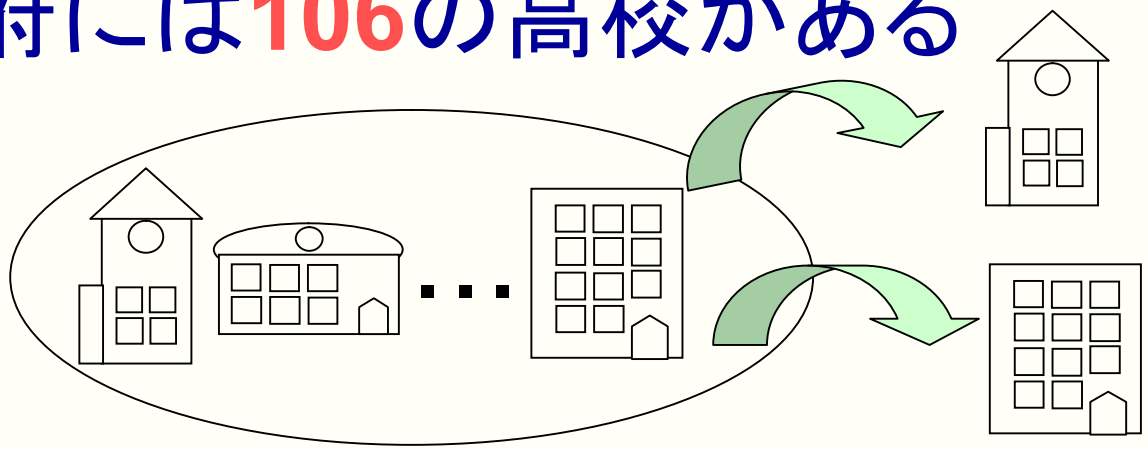


ターゲット集団

サプリメントの購入調査の例

- 学校をクラスターとすると

京都府には106の高校がある

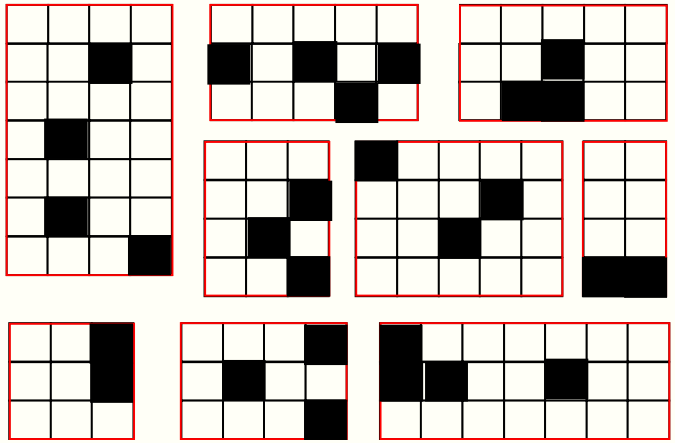


- 学校（クラスター）をランダムサンプリング
- 選ばれた学校の全員を調査

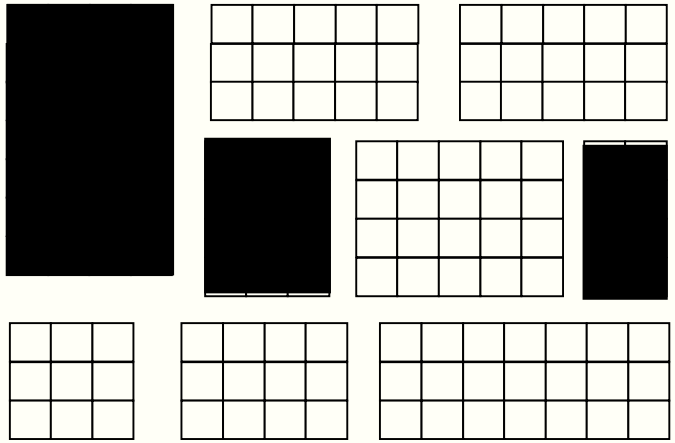
クラスターサンプリングでの推定

- サンプルリングの対象はクラスター
  - サンプルとなったクラスターが
    全クラスターの代表
  - ばらつく要因は個人ではなくクラスター

層化とクラスターの違い



層化サンプリング



クラスター
サンプリング

層化とクラスター

- グループの役割は異なる
- 層化サンプリング
  - 推定値の精度が上がる
- クラスターサンプリング
  - （通常）推定値の精度は下がる

多段階サンプリング
multistage sampling

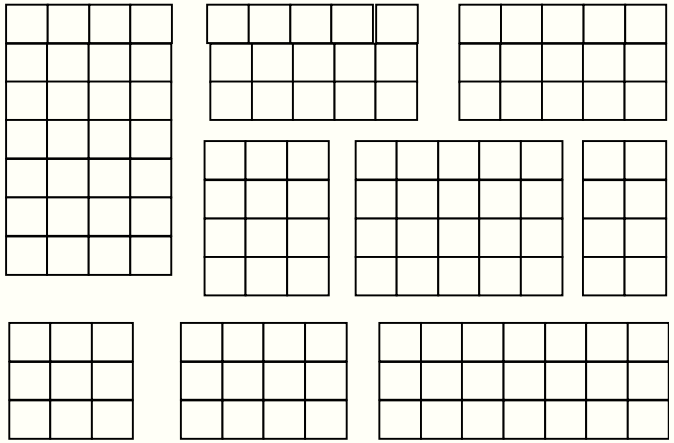
- クラスターの中にクラスターがある

  例） 病院-診療科

  学校-学級-生徒

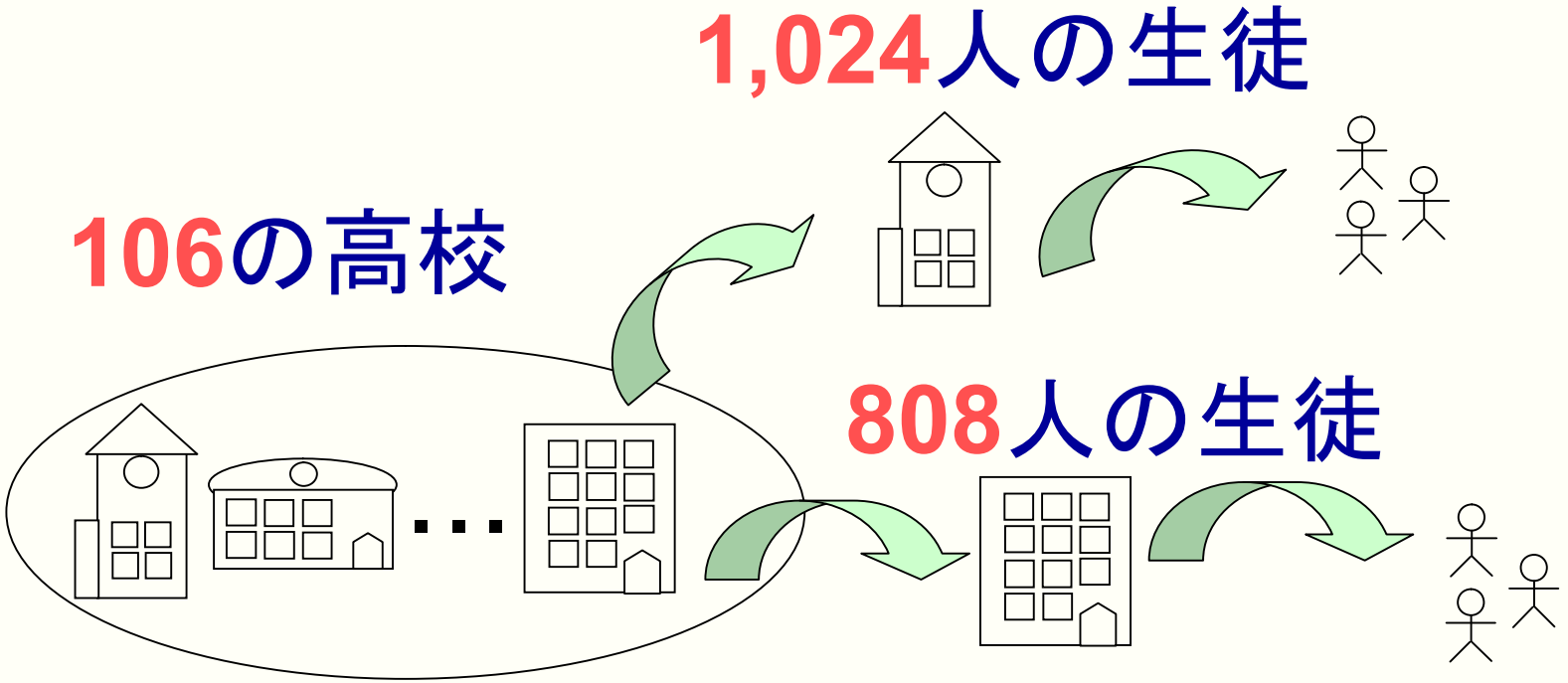
- これもクラスターサンプリングと呼ばれることもある

2段階サンプリングのイメージ



ターゲット集団

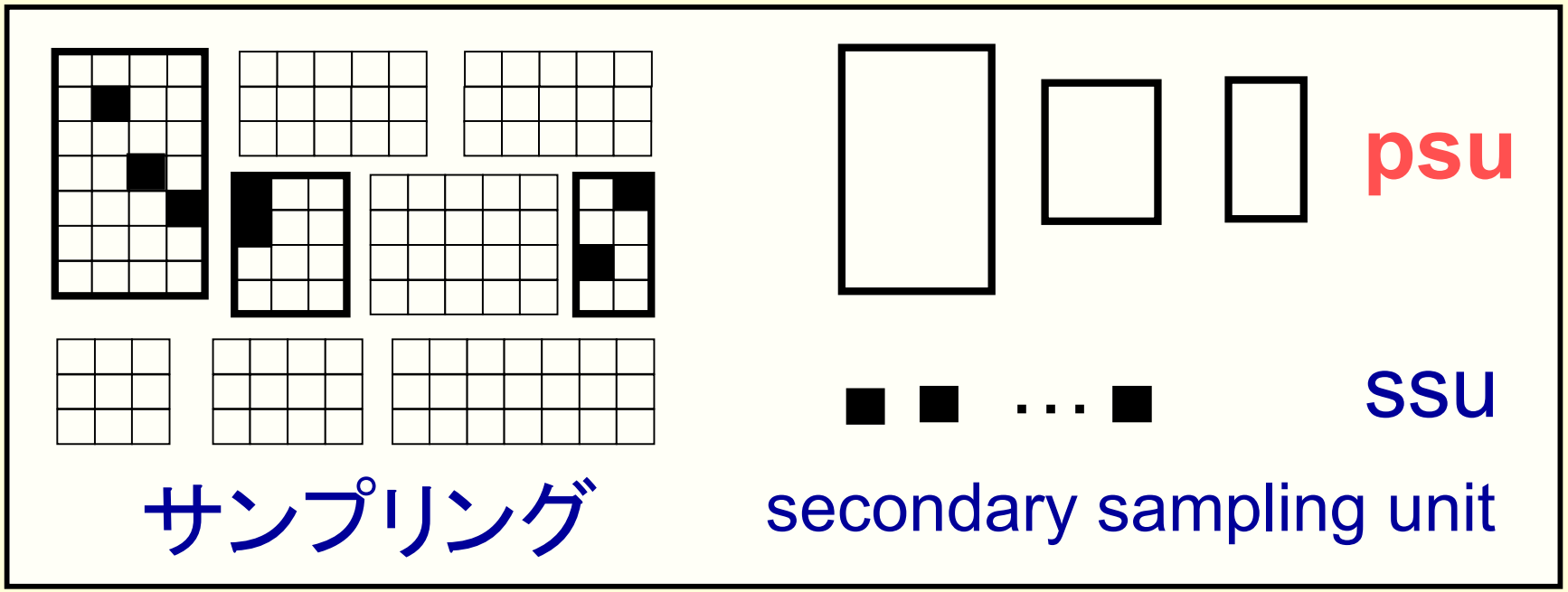
サプリメントの購入調査の例



- 学校（クラスター）をランダムサンプリング
- 選ばれた学校の生徒をサンプリング

一次サンプリングユニット
primary sampling unit (psu)

- ターゲット集団から直接サンプルされるグループ



多段階サンプリングでの推定

- 重みを算出
  - ターゲット集団のクラスターの数と選択されたクラスター内の人数がわかっている必要あり
- 重み付平均を計算

推定値の精度

- 推定値の精度は psu と ssu に依存
  - えらく複雑
- psuのみで計算可能な近似がある
  - pusが何かを把握することが重要

Q. アイデアを募集

- 層化サンプリング
  - ➢推定値の精度が上がる
- クラスター、多段階サンプリング
  - ➢実現可能性が高い
- なにかいいアイデアは？

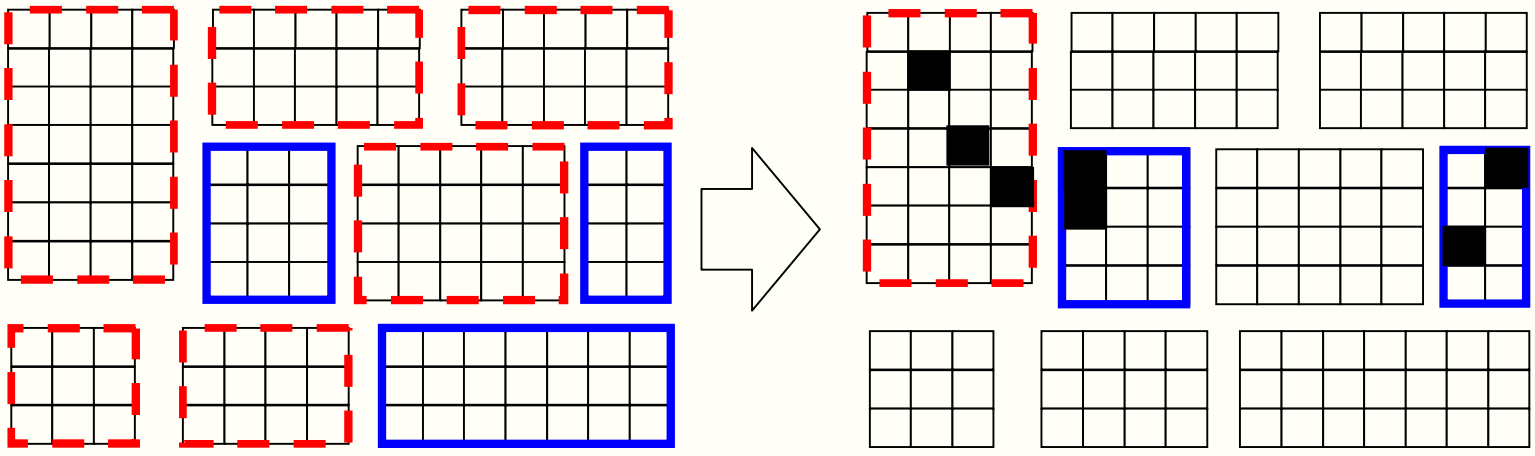
層化多段階サンプリング
stratified multistage sampling

- 現実的な多段階サンプリング

＋

- 精度をよくする層化サンプリング

層化2段階サンプリングのイメージ



ターゲット集団　　サンプリング

Q. 層化2段階サンプリング

- 層化2段階サンプリングで、
  平均値と標準誤差を計算するために
  必要な情報(データ)は？
  - どの層にいくつのクラスターがあるか
  - 各クラスターの対象者の数

Q. サプリメントの購入調査

- 京都府の例

学校区分が層で、学校がクラスターの場合

学校区分	学校数
国立	1
公立	64
私立	41
計	106

各学校の生徒数

層化多段階サンプリングでの推定

- 層とクラスターに関する重みを算出
  - 各層のクラスターの数と
    各クラスターの人数から
- 重み付平均を計算

平均値と標準誤差を推定するために

- サンプリングにより得られた調査の回答
- サンプリング法に基づく重みがいくらか
  - 一人が何人分を代表しているか
- psuが何か

Q. 疫学研究とサンプリング

- 疫学の授業で習った研究デザインのサンプリングは？

疫学研究とサンプリング

- 今まで習った疫学研究のデザイン
  - 2つのグループを偏りなく比較

- 今日のサンプリングの話
  - ターゲット集団の特徴を偏りなく推定
  - ターゲット集団が有限

まとめ

- サンプリングによって、
  ある精度で集団の特徴を知ることができる

- サンプリングの種類
  - ➢シンプルランダムサンプリング
  - ➢層化サンプリング
  - ➢クラスター、多段階サンプリング

まとめ

- 調査回答の平均値と標準誤差は、
  回答と重みとpsuから

補足　重み付平均

層　$h=1, 2, \ldots, H$

クラスター　$i=1, 2, \ldots, n_h$

個人　$j=1, 2, \ldots, m_{h\cdot i}$

第h層、第iクラスター、第m番目の人の

重み：w_{hij}

回答：y_{hij}

補足　重み付平均

重み付平均

$$\frac{\sum_h \sum_i \sum_j w_{hij} y_{hij}}{\sum_h \sum_i \sum_j w_{hij}}$$