データサイエンスの基礎知識

医療保健学部　医療情報学科　横堀滉弥

ターゲット集団（母集団）　～興味の対象となっている集団～

①ターゲット集団の全員からデータを取る⇒**全数調査**

メリット：正確に把握できる

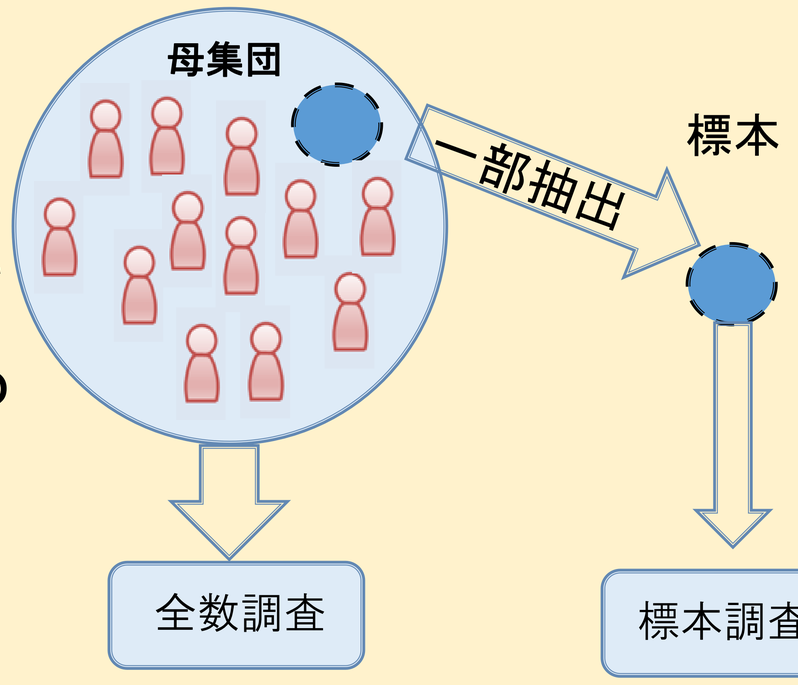
デメリット：ターゲット集団が大きいと実施が困難

②ターゲット集団の一部（**標本**）からデータを取ってもし全員調べたらどのような結果がでるか推測する

⇒**標本調査**

メリット：実施しやすい

デメリット：正確？



☆ランダムサンプリングの背景☆

標本調査をする際に、ターゲット集団の様子を正しく推測するためには、標本がターゲット集団の様子を反映している必要がある。そのためサンプリングの方法としてランダムサンプリングという方法が用いられる。

ランダムサンプリングとは

標本を作るために、ターゲット集団から対象をランダムに抽出すること。

どうして一般化可能性の議論を行うのか？

ランダムサンプリング以外の方法で標本が作られるとき、その標本がターゲット集団の偏った一部になってしまう。

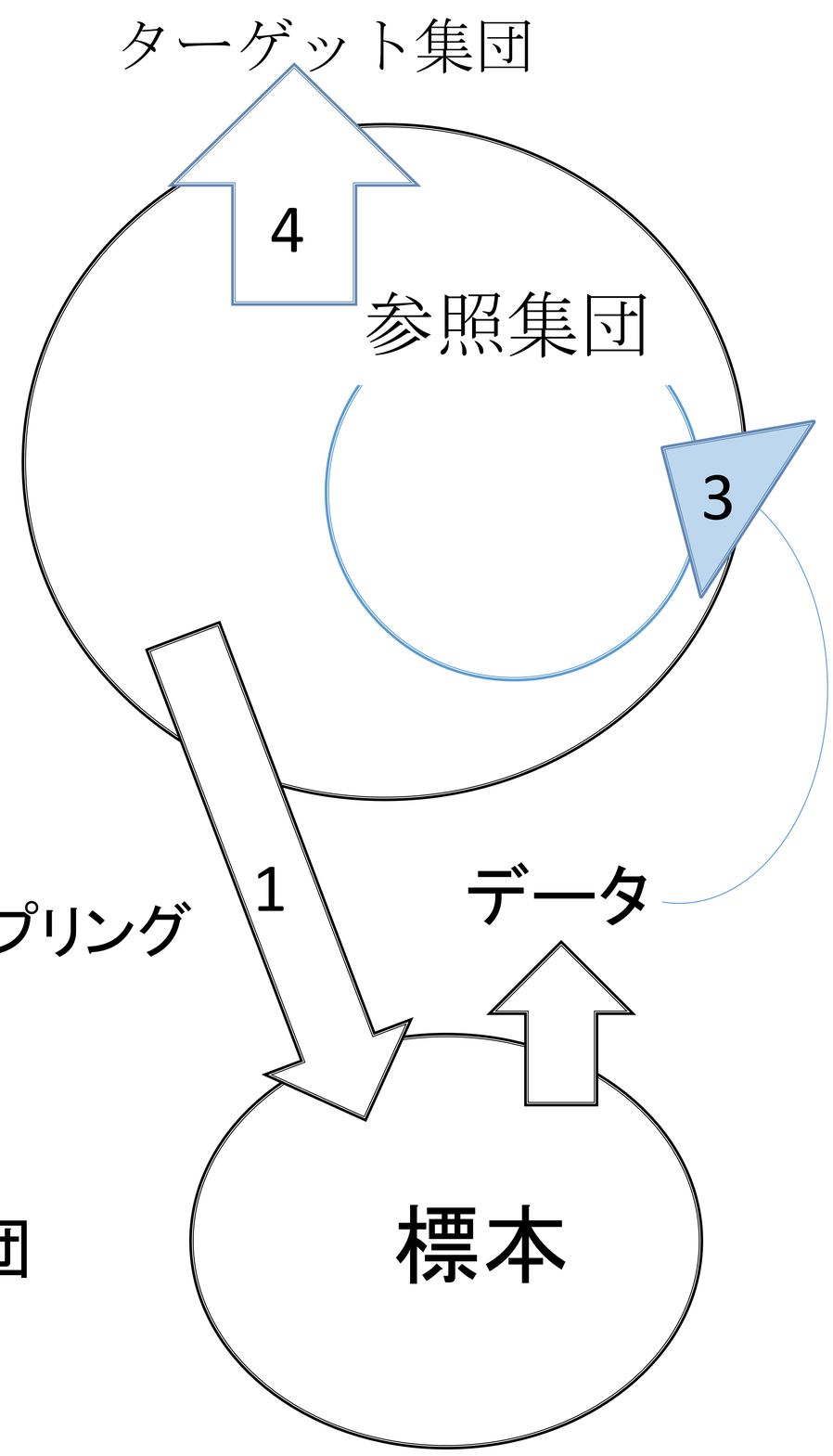
そうするとこういったことが起こる。

1.直接ターゲット集団の様子を推測することはできない

2.標本と同様に偏った標本より大きい集団（参照集団）を想定する

3.参照集団を推測

4.ターゲット集団の様子を議論する



・一般可能性の議論

参照集団への推測結果を用いてターゲット集団の様子を議論すること

推測統計（手法）

標本調査で必要となる手法。標本より大きな集団（ターゲット集団、参照集団）の様子を表現する。アプローチとして、**推定**と**検定**がある。

推定とは？

参照集団全員を調べたときに分かる値を標本のデータを用いて**具体的な数値**として表現する方法。

具体的な数値として

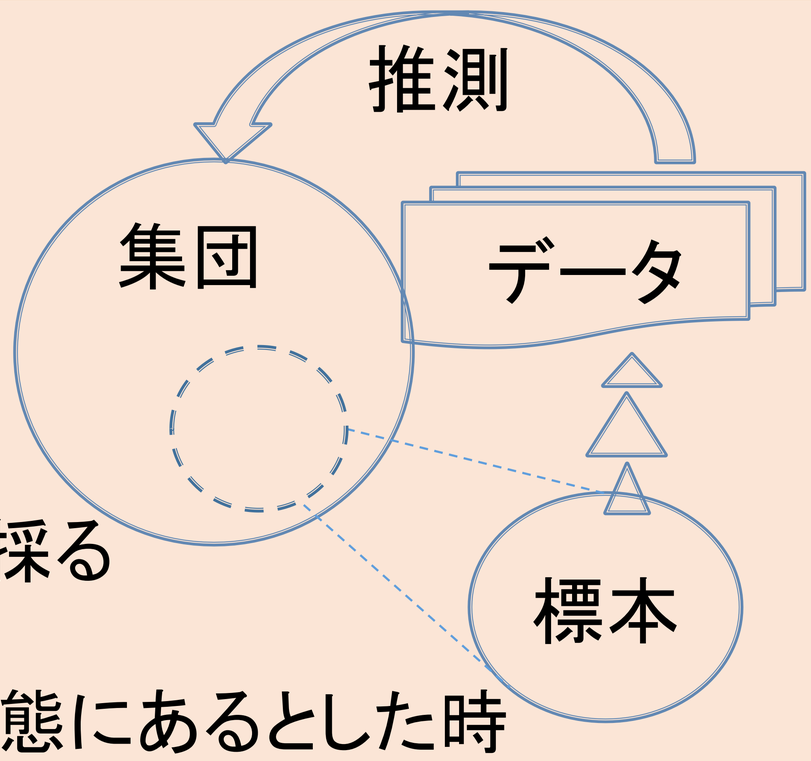
一つの数値で表現⇒**点推定**　区間で表現⇒**区間推定**

検定とは？

参照集団全員を調べたい時にわかることについて仮説を立てその仮説が誤っているかどうか標本のデータを用いて判断する方法。

集団に対する仮説が、間違っているのかをデータを用いて判断する。

検定の考え方



背理法の手順にしたがっている

①帰無仮説を立てる！

集団全員を調べた時の様子に関して仮説を立てる。

②集団を観察する

集団からランダムサンプリングした標本からデータを採る

③データと帰無仮説が矛盾していないかを調べる

集団全員を調べた時の様子が、帰無仮説のような状態にあるとした時標本から採られた手元のデータが出現する確率「**p値**」を計算

④帰無仮説について判定を行う

＜p値が有意水準より大きい場合＞

帰無仮説の下でこのデータが出現してもおかしくない⇒データと帰無仮説に矛盾なし

⇒帰無仮説の判断を保留する

＜p値が有意水準より小さい場合＞

帰無仮説の下でこのデータが出現するのは信じがたい⇒データと帰無仮説に矛盾あり

⇒帰無仮説を否定する

有意水準

検定において帰無仮説を設定したときにその帰無仮説を棄却する基準となる確率のことです。α（アルファ）で表され、5%（0.05）や1%（0.01）といった値がよく使われます。