実験データー解析概論

— 統計学に基づく「よりよい推論」のために —

三中 信宏

MINAKA Nobuhiro

独立行政法人 農業環境技術研究所 生態系計測研究領域 上席研究員［生物統計学］

東京大学大学院 農学生命科学研究科 生物・環境工学専攻 教授［生態系計測学］

東京農業大学大学院 農学研究科 客員教授［応用昆虫学］

mailto:minaka@affrc.go.jp （メール）

http://twitter.com/leeswijzer/ （ツイッター）

http://cse.niaes.affrc.go.jp/minaka/ （ウェブサイト）

http://d.hatena.ne.jp/leeswijzer/ （ブログ）

2012 年度・東京農業大学生物応用化学科（毎週金曜第 2 限）

三中信宏〈実験データー解析概論〉：2012 年 4 月～ 2012 年 9 月

東京農業大学世田谷キャンパス 431 番教室

母集団からのサンプリング

標本から計算された「統計量」はどれくらいのバラツキを伴っているのか？

リンゴ 10 個（標本）

統計量（サイズ，糖度，. . . ）

サンプリング

推定・検定

リンゴ園（母集団）

統計量に伴う「誤差」を標本データから推定すればよい

母集団からのサンプリング

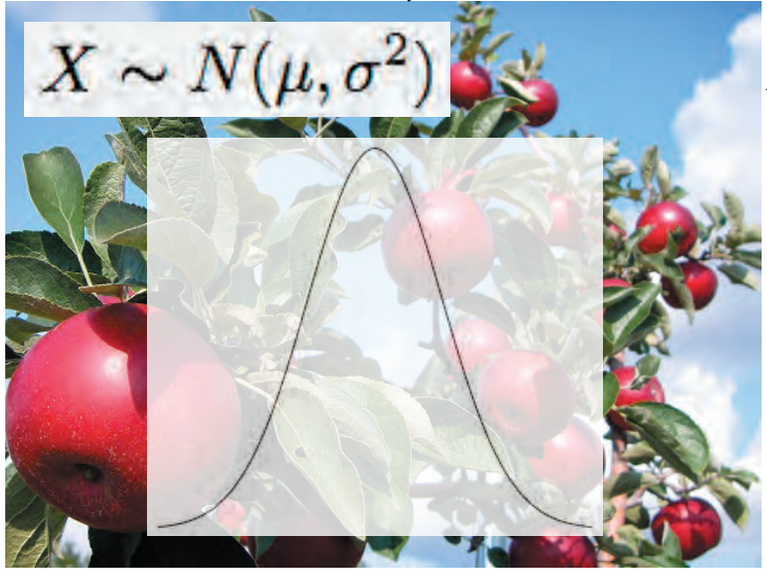
母集団が正規分布するならば，標本平均もまた正規分布をする

標本

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

サンプリング

推定・検定



母集団

$$var(\overline{X}) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n(n-1)}$$

標本平均の分散推定値は標本データそれ自体から計算できる

母集団からのサンプリング

なぜ苦しまねばならないのか？

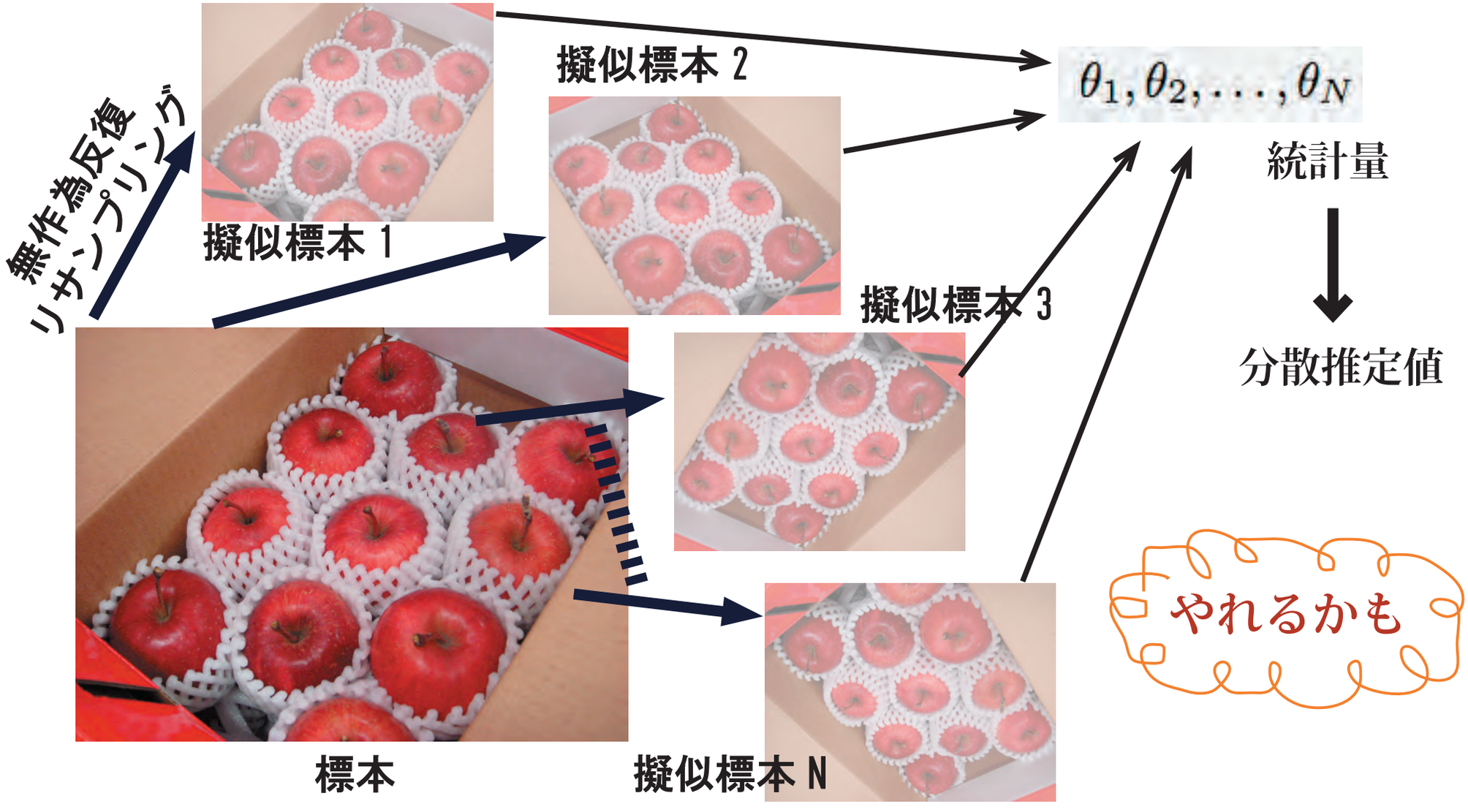
統計量の誤差の推定値は，正規分布に従う母集団のもとでの単純な統計量の場合を除いては，一般に解析的に求めることはできない．したがって，そのような制約を緩める数値的方法が必要になる．

リサンプリング統計手法（resampling methods）

1) ブーツストラップ（bootstrap）
2) ジャックナイフ（jackknihe）
3) モンテカルロ（Monte Carlo）

データからの“らくらく”リサンプリング

標本を仮想的母集団とみなしてしまおう



ブーツストラップ

重複を許して無作為同数リサンプリングを反復する

データ = 仮想母集団

$=$

$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ (元)

1

$X_1, X_1, X_3, X_4, X_6, X_7, X_7, X_8, X_9, X_{10}$

2

$X_2, X_3, X_5, X_5, X_6, X_7, X_8, X_8, X_{10}, X_{10}$

N

$X_1, X_3, X_3, X_4, X_5, X_7, X_7, X_8, X_9, X_{10}$

同数の標本抽出

$\theta_1, \theta_2, \ldots, \theta_N$

各反復から計算された統計量の集合

$$var(\hat{\theta}_B) = \frac{\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}{(N-1)}$$

分散のブーツストラップ推定値

ジャックナイフ

重複を許さず無作為削除リサンプリングを反復する

データ = 仮想母集団

$=$

$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ (元)

1

2

N

一つ削って標本抽出

$X_1, \overset{v}{X_3}, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$

$X_1, X_2, X_3, \overset{v}{X_4}, X_6, X_7, X_8, X_9, X_{10}$

$\overset{v}{X_2}, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$

$\theta_1, \theta_2, \ldots, \theta_N$

各反復から計算された統計量の集合

$$var(\hat{\theta}_J) = \frac{\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}{(N-1)}$$

分散のジャックナイフ推定値

パラメトリック・ブーツストラップ
~~モンテ・カルロ~~

パラメータ推定値に基づくデータ生成シミュレーション

データ = 仮想母集団

$=$ $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ (元)

推定値 $\hat{\mu}, \hat{\sigma}^2$

$N(\hat{\mu}, \hat{\sigma}^2)$

モデル

擬似データ生成

$\theta_1, \theta_2, \ldots, \theta_N$

各反復から計算された統計量の集合

$$var(\hat{\theta}_B) = \frac{\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}{N(N-1)}$$

分散のシミュレーション推定値

母集団が従う統計モデル（確率分布や線形モデル）を仮定し，そのパラメータを元データから推定する．そのパラメータ推定値をもつモデルから擬似データの生成シミュレーションを反復試行し，その結果から統計量の分散を推定する．