データ解析のための統計モデリング入門

1．データを理解するために統計モデルを作る

新納浩幸

統計モデルの特徴

- 観測によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり、これはデータに見られるばらつきを表現する手段である
- データとモデルを対応づける手続きが準備されていて、モデルがデータにどれくらいよくあてはまっているかを定量的に評価できる

統計モデルが有効な理由

確率分布を使うことで
「ばらつき」「欠損」などを
うまく表現できる

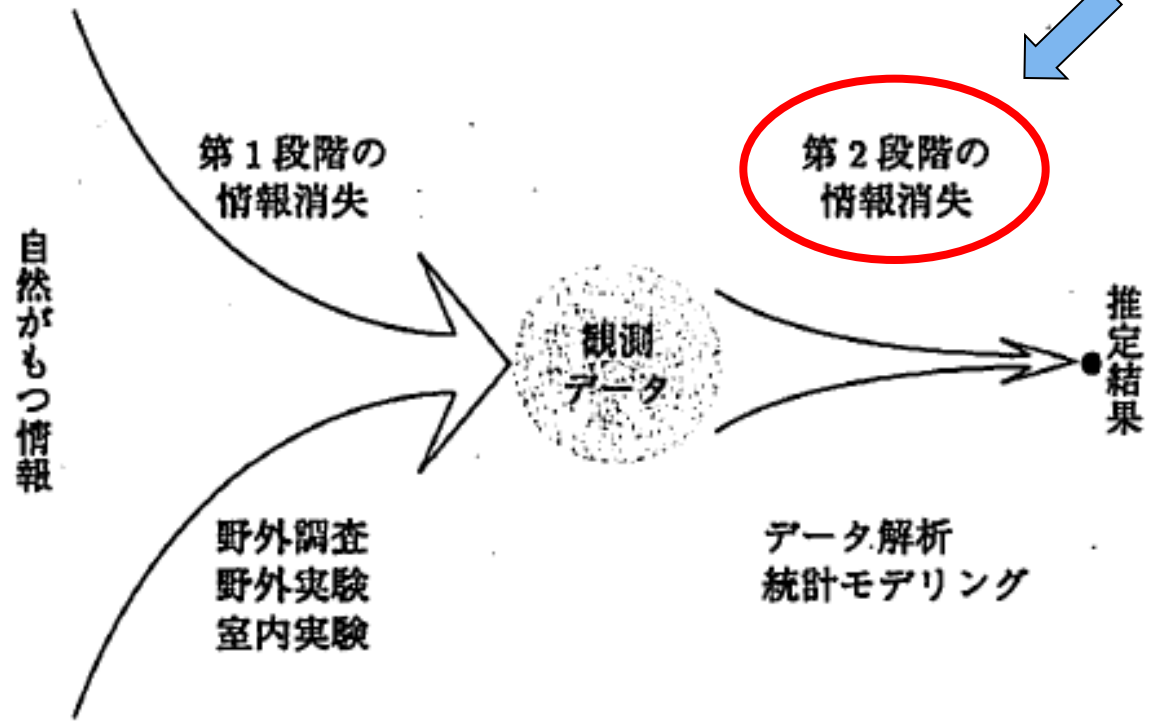


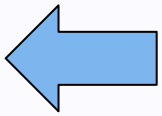
図 1.1　自然科学における２段階の情報の消失．この本ではもっぱら第２段階だけをあつかう．

ブラックボックスな統計解析

- 「ゆーい差」が出るまで検定方法をひたすら取り替える
- データ中の観測値どうしの割り算によって新しい「指標」をでっちあげる

・・・

- 論文中でデータを示すときは何でも検定して P 値をつける、P 値が小さいほど自分の主張はただしい



手法の中身を理解しないで、データをこねくり回して、勝手な解釈するな・・・っということ

本書の内容

一般化線形モデル（GLM）と呼ばれるクラスの統計モデル、
そしてそのベイズ化によるモデルの拡張・・・・が中心

線形モデル（LM）・・・データのばらつきが等分散正規分布を仮定
一般化線形モデル（GLM）・・・上記の仮定を緩める

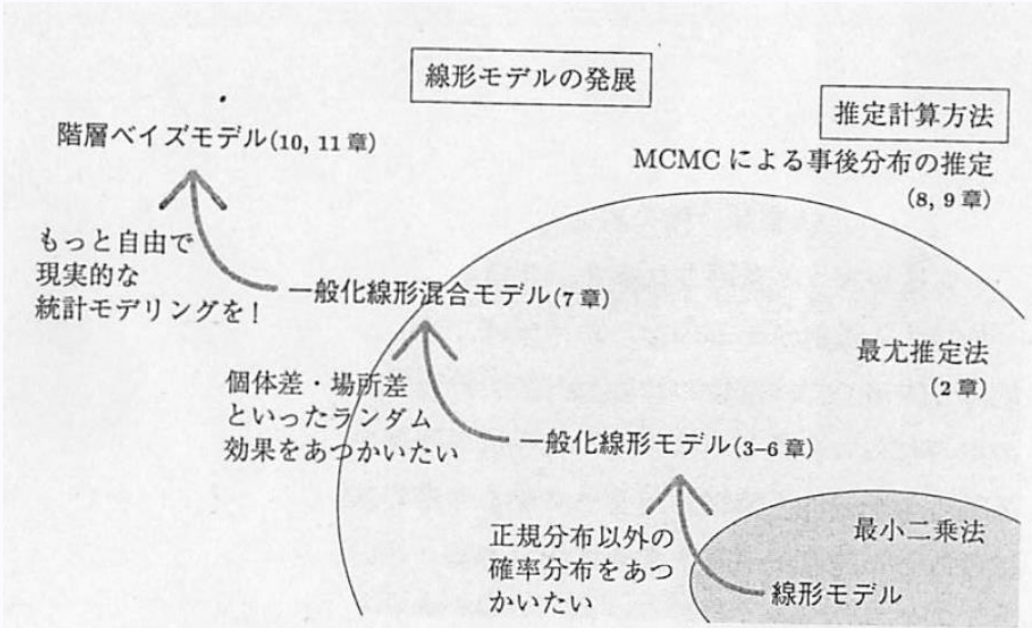


図 1.2　線形モデルを発展させる説明のプラン．まずポアソン分布や二項分布を使った一般化線形モデル（GLM）を導入し，それを現実的なデータ解析に使えるように階層ベイズモデル化する．