データ解析のための
統計モデリング入門

一般線形化モデル・階層ベイズモデル・MCMC

この本のねらいと想定している読者

- 想定している読者

  ➔「数理モデルで現象を表現・説明する」基礎訓練を受けていない人

- あつかう内容

  ➔ 一般化線形モデル (GLM)の基礎と発展

  個体数の変動、空間分布、環境に対する個体の応答などの内容のデータを使用します。

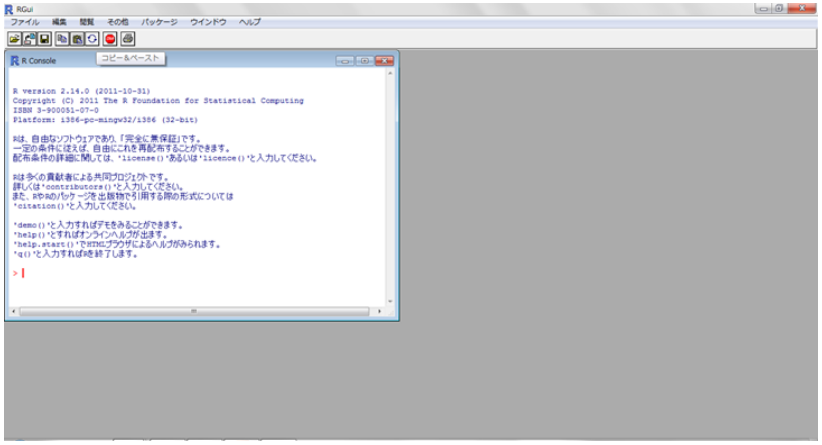
使用する統計ソフトウエアとサポートwebサイト

・R

➔ データの図示、要約、統計モデルのあてはめ

・WinBUGS (9章～)

➔ 階層ベイズモデルのパラメータ推定



Rの画面

Rのいいところ

- Freeのソフトウェア
- ソースコードも完全公開
- 作図機能がある
- プログラミングができる
- 機能拡張が容易

などなど、とっても便利！

この本のサポートサイト

http://goo.gl/Ufq2

（例題のコードなどが掲載されています。）

第1章
データを理解するために
統計モデルを作る

1.1　統計モデル：なぜ「統計」な「モデル」？

○統計モデル（statistical model）

- 観測によってデータ化された現象を説明
- 確率分布が基本的部品

  データの「ばらつき」「欠損」をうまく表現できる！
- モデルがデータにどれくらいよくあてまっているか、定量的に評価

  モデルの信用できなさやモデルによる予測精度の限界も示せる

解析者の意図を明確に表現した統計モデル

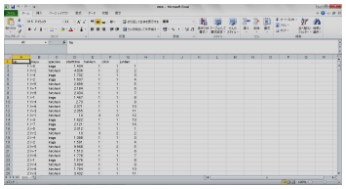
→　他の研究者とのアイデア共有が簡単に！

複雑な情報をわかりやすく

自然がもつ情報

第1段階の
情報圧縮

野外調査
野外実験
室内実験



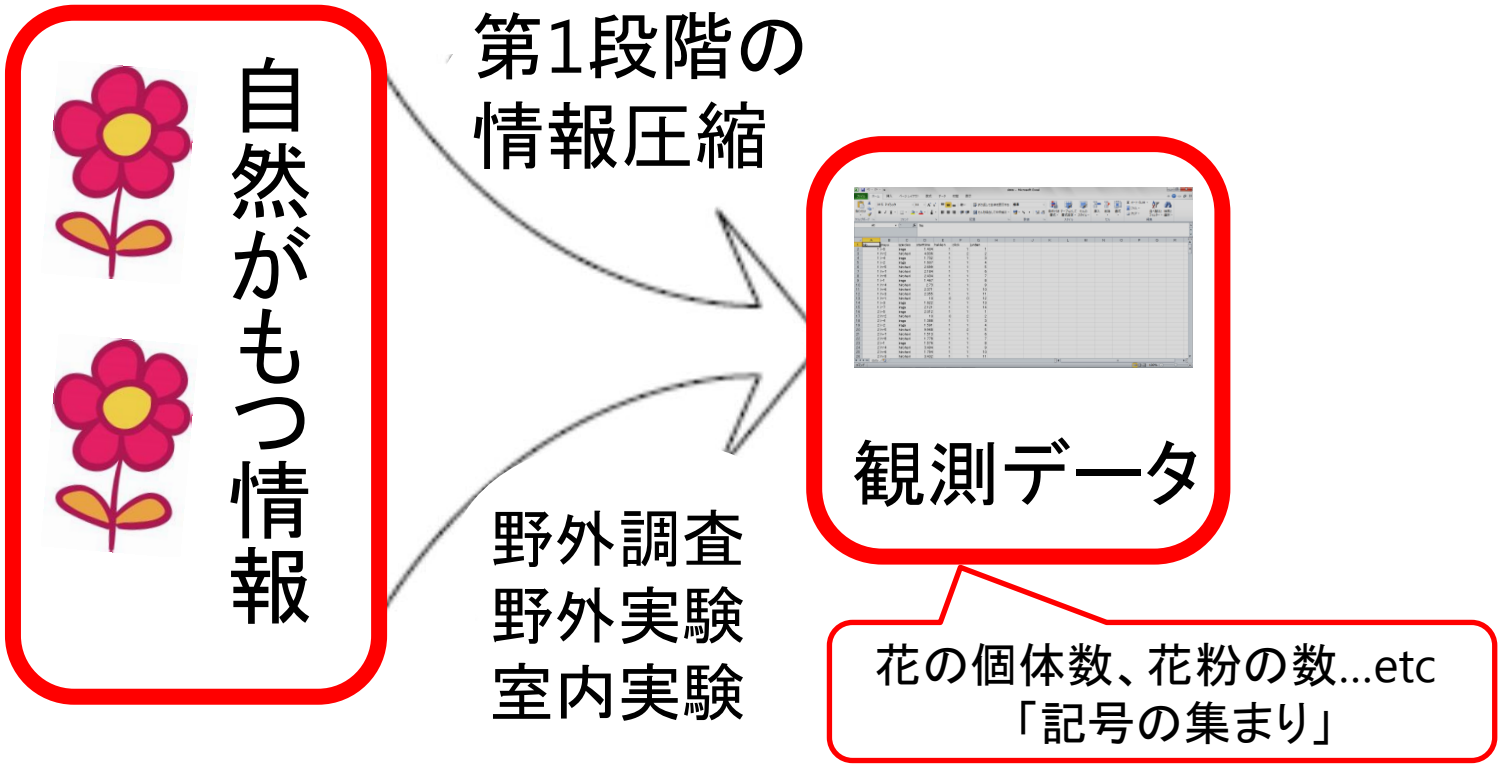
観測データ

第2段階の
情報圧縮

データ解析
統計モデリング

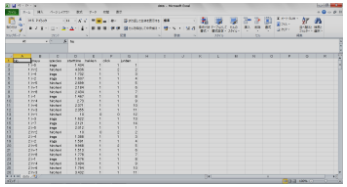
推定結果

第1段階の情報圧縮 →この本では具体的にあつかいません



目的に応じて、観測・実験といった手段で
対象からの情報をとりだす

第2段階の情報圧縮 →この部分を集中的に扱う



観測データ

記号の集まりでよくわからない

第2段階の
情報圧縮

データ解析
統計モデリング

推定結果

統計モデルのあてはめによって、
情報を整理し、よりわかりやすくする

実験設計を考えるときは、統計モデリングを
その後やるのを頭に入れて設計したほうがいい！

1.2 「ブラックボックスな統計解析の悪夢」

ブラックボックス統計学

＝「理解しないまま、ソフトウエアを使う」

- 「ゆーい差」が出るまで検定方法をとりかえる
- R^2値は「説明力」なので、ひたすら1に近ければよい
- 「検定を何度もやっているので多重比較だ」と文句をつけられれば、
  何でもかんでも多重検定法による補正をやればよい　などなど

理屈にあわないのに、データを入れると、
それらしい出力が得られてしまうため起こる！

解析者がデータをよく見て、目的に沿いつつ、構造に合致した統計モデルを構築するのが大事

1.3　一般化線形モデルの導入とそのベイズ的な拡張

階層ベイズモデル

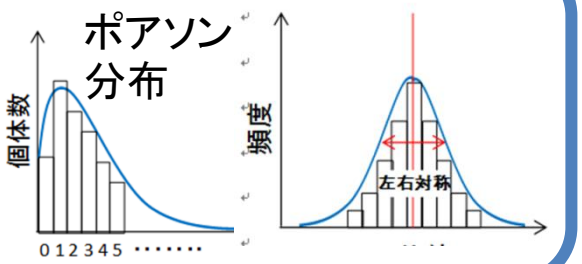
もっと自由にモデリング！

一般化線形混合モデル(GLMM)

ランダム効果（個体差、場所差）も扱える

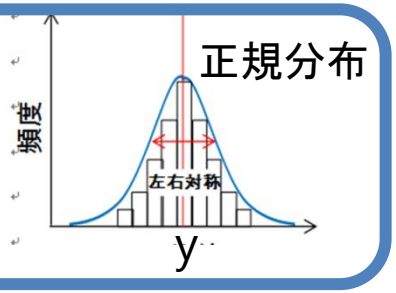
一般化線形モデル（GLM）

さまざまな確率分布を扱える



線形モデル（LM）

等分散正規分布を仮定



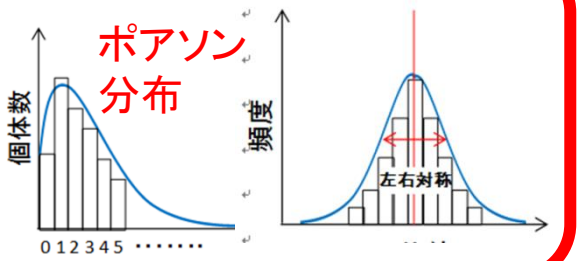
1.3.1　各章の内容（2章）

階層ベイズモデル

一般化線形混合モデル(GLMM)

一般化線形モデル（GLM）

さまざまな確率分布を扱える



線形モデル（LM）

確率分布について学ぶ（ポアソン分布が中心）

カウントデータをうまく表現できる確率分布

最尤（さいゆう）推定の考え方を学ぶ

統計モデルにデータをあてはめる→パラメーターを推定値

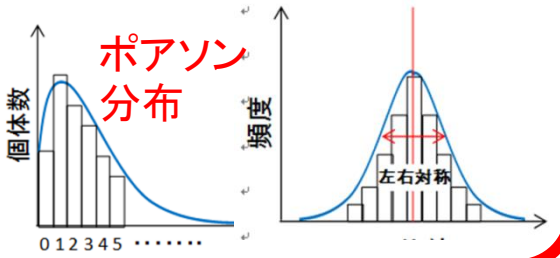
1.3.1 各章の内容（3章）

階層ベイズモデル

一般化線形混合モデル(GLMM)

一般化線形モデル（GLM）

さまざまな確率分布を扱える



線形モデル（LM）

GLMの詳細ついて学ぶ

ポアソン分布＋リンク関数＋線形予測子→統計モデル

1.3.1　各章の内容（4章・5章）

複数のモデルの良し悪しを比較する方法を学ぶ

- AIC（赤池情報量規準）（4章）
  - モデルのあてはまり度を表す統計量。予測の良いモデルをモデル選択する。
  - AICは小さい程、そのモデルがデータを把握できている

ゆうど
- 尤度比検定（5章）
  - モデル間の最大対数尤度（観測されたデータへの当てはまりの良さ）を比較する検定方法

モデル選択（4章）と検定（5章）の違いに注意

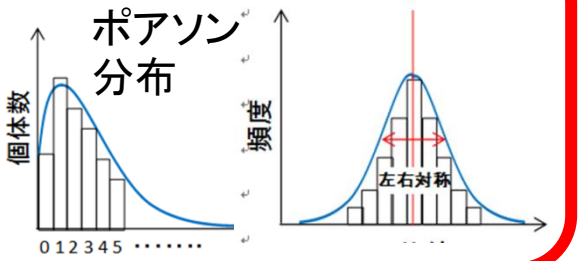
1.3.1　各章の内容（6章）

階層ベイズモデル

一般化線形混合モデル(GLMM)

一般化線形モデル（GLM）

さまざまな確率分布を扱える



線形モデル（LM）

二項分布、正規分布、ガンマ分布を用いたGLM

これを部品として、ロジスティック回帰を行う

1.3.1　各章の内容（7章）

固定効果・ランダム効果をモデルに組み込んだ、
GLMMの考え方を勉強する

1.3.1　各章の内容（8章～）

階層ベイズモデル

もっと自由にモデリング！

一般化線形混合モデル(GLMM)

一般化線形モデル (GLM)

線形モデル (LM)

より複雑なモデルになるので、最尤推定が難しい・・

MCMC（マルコフ連鎖モンテカルロ法）で推定

↓

ベイズ統計モデルで考えたら便利そう！

1.4 この本に登場する訳語・記号・記法について

○定まっていない訳語

- 逸脱度（devidance）
- 残差逸脱度（residual devidance）
- 最大逸脱度（null devidance）

あてはまりの悪さ

- 固定効果（fixed effects）
- ランダム効果（random effects）
- リンク関数（link function）

たとえば、GLMの式は…

リンク関数

係数(β_2、β_3)

線形予測子

$$f(y_i) = \beta_1 + \beta_2 x_i + \beta_3 f_i + \cdots$$

切片(β_1)

説明変数(x, f_i)

y_i：応答変数(y)の確率分布の平均

確率分布から得られるリンク関数
→線形予測子を構築
→パラメーターを推定

○この本での記号や数式の記法

- 積の記号やかっこは省略

$$\beta_2 \times x_i = \beta_2 x_i \qquad \log(x) = \log x$$

- 指数関数・対数関数について

$$\exp x = e^x \qquad \log x \leftarrow e\text{を底とする関数}$$

- 架空植物の個体などを表す添え字は i を使う

$i \in \{1, 2, 3, \ldots 50\}$ = iは個体番号1～50までをとる
（$\in$：属する）

○記号や数式の記法

- y_iを扱う際に、この量の予測でyを使うことがある。
  y_i の中のどれでもいい→y_*と表記

- 和・積の記号の下に*i*があれば、
  すべての個体についての和・積を示す

$$\left(\begin{array}{l} \sum_{i \in \{1, 2, 3, \ldots 50\}} \log L_i^* \rightarrow \sum_i \log L_i^* \\ \\ \overset{\text{パイ}}{\prod_{i \in \{1, 2, 3, \ldots 50\}}} L_i^* \rightarrow \prod_i L_i^* \end{array} \right)$$

- ≈（近似）は使わず、=（等号）で表記

○この本での記号や数式の記法

- 比例をあらわす二項演算子 $\propto$（比例）を使う

$$p = q \times (\text{定数}) \rightarrow p \propto q$$

- 確率変数がある確率分布にしたがうとき、
  ～記号で表現する

$$y_i \sim (\text{平均 } \lambda_i \text{のポアソン分布})$$

（λ：ラムダ）

- 事象AとBが生起する同時確率 $p(A, B)$
  Bの条件下でAが生起する条件付き確率 $p(A \mid B)$

○この本での記号や数式の記法

・x_i 全体の集合
$\{ x_1, x_2 \ldots\ldots x_N \} \rightarrow \{ x_i \}$ と表記

・集合・ベクトル・行列などは太字表記しない
ただし、$\{ y_i \}, \{ x_i \} \rightarrow \mathbf{Y}, \mathbf{X}$ と表すこともある

$$\mathbf{X} = \{ x_i \} = \{ x_1, x_2 \ldots\ldots x_N \}$$

1.5 第1章でのまとめ

データ解析において、、

「統計モデルを理解しながら使う」
ことは重要！！！