2004.08.28

日本生態学会釧路大会の自由集会

データ解析で出会う統計的問題— 多重検定と多重比較をめぐって

統計ソフトウェア R でやってみる多重比較

— あなたの研究に「検定」が必要ですか? —

http://hosho.ees.hokudai.ac.jp/~kubo/ce/2004/

講釈: 久保拓弥　kubo@ees.hokudai.ac.jp

今日のハナシ: 多重検定から「逃げる」

多重検定を使わねばならぬ状況はある— しかしもっと簡単な「モデル選択」で十分な場合も多々ある

1. **まえおき**
   誰がための検定，多重な比較，モデル選択，そして R
2. **架空の例で考えてみる多重比較なモデル選択**
   一番簡単な例題で問題の解きかたを検討する
3. **R で強める多重比較なモデル選択**
   R プログラミングで難しい問題も解決できる

まずは……「誰がための検定」

統計学的検定とは何か?

1. 帰無仮説という「だめ仮説」を作る (検定の非対称性)
2. 「だめ仮説」を間違って捨てる (第 I 種の過誤) 確率を計算

統計学的検定から得られる結果の特徴

つよみ: 帰無仮説を誤って棄却する危険性 (p 値) は「保証」されている

よわみ: 帰無仮説が棄却できない $\Leftrightarrow$ 何も言えない

- cf. 検定力 (power)

あなたの研究で結論を述べるためには，このようにして計算された p 値が必要ですか?

検定が必要なヒト，必要でないヒト

ここで研究者を便宜的に二種類に分類できる

p 値が必要である (「かたぎ」な世界)

- 第 I 種の過誤の回避について，きちんとした手続きしたい
- 実験計画法を正しく使って必要とされる標本数を事前に根拠にもとづいて算定している

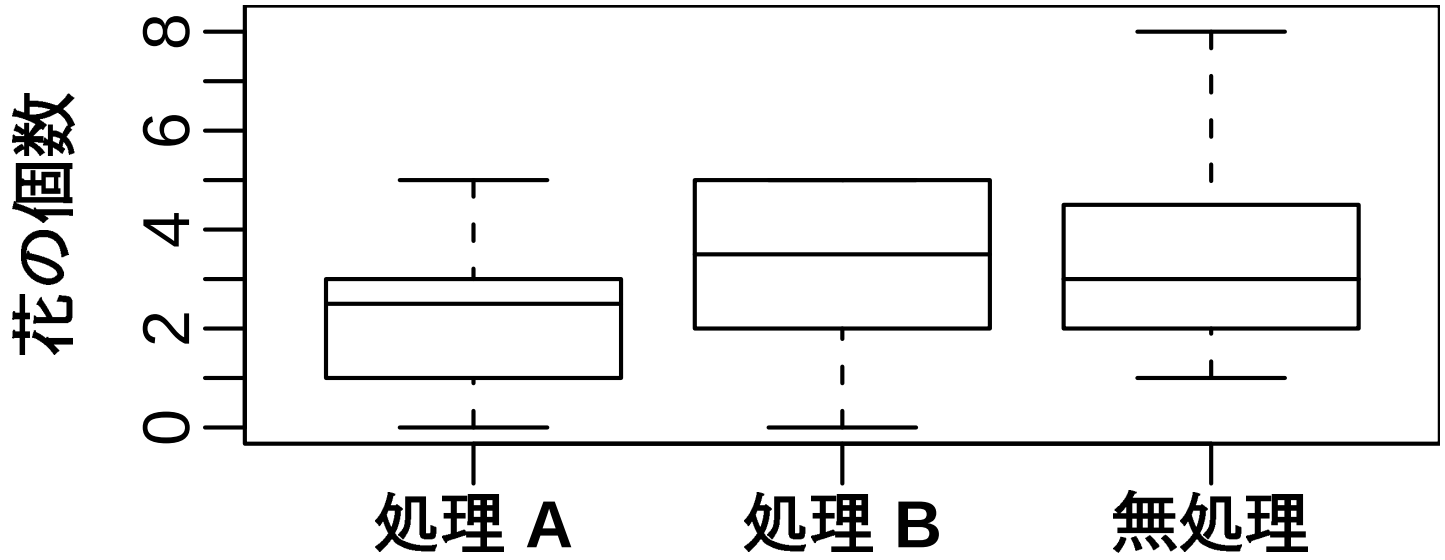
p 値，必要ないかも (「やくざ」な世界)

- 「第 I 種の過誤だけが重要」というわけでもないなぁ
- 「実験計画法って何?」

本日は (というかいつもだけど)，ここから先は「やくざ」な連中のためのハナシをします

じつは，これをやりたいだけなんでは?

もし何かの植物に関するこういう観測データがあった場合，



- **処理** A **は処理** B **とも無処理** (control) **とも違っていた**
- **しかし処理** B **は無処理と違っていなかった**

これを言うためだけに，帰無仮説や第 I **種の過誤の確率 (いわゆる** p **値) が必要なんだろうか? (**p **値だけを重視すべき理由を持たないときに)**

2004.08.28

統計学的な「モデル選択」で対応できそう

モデル選択って何?

1. モデル選択を適用したい範囲 (標本集団) を決める
2. 適用する確率論的モデル (統計モデル) を列挙する
3. それぞれの統計モデルのパラメーターの最尤推定値を得る
4. モデル選択基準を計算する
5. モデル選択基準が最良のものを採用する

科学で「客観的」な ($\neq$ 絶対正しい) 何ごとかを述べる手段としての要件は満たしている

モデル選択基準: (いろいろあるんだけど) Akaike's Information Criteria (**赤池の情報量基準**)

ここでは **AIC** 使ってみる　これが小さいほど「良い」モデル

$$\mathrm{AIC} = -2(\textbf{最大化対数尤度}) + 2\,(\textbf{パラメーター数})$$

あてはまりぐあい (良い)　モデルの複雑さ (悪い)

これ使いましょう: 統計ソフトウェア R

http://www.r-project.org/

- いろいろな OS で使える free software
- 使いたい機能が充実している
- 作図機能も強力
- S 言語によるプログラミング可能
- よい教科書が出版されつつある
  - 「The R Book」 岡田昌史編 (2004)
  - “Modern Applied Statistics With S” Venables & Ripley (2002)
  - “Introductory Statistics with R” P. Dalgaard (2002)
  - ネット上のあちこち (これがもっとも重要)

架空の例で考えてみる
多重比較なモデル選択

一番簡単な例題で問題の解きかたを検討する

[重要な技法] 架空の数値例を生成し統計的手法を適用する

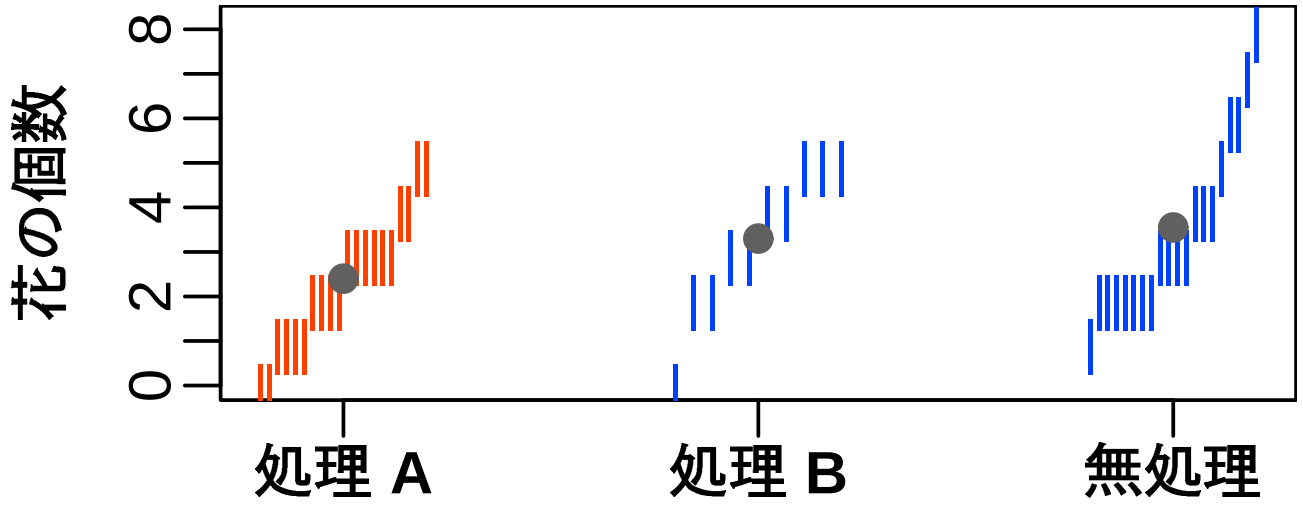
1. 母集団を自分で決める
2. R の乱数生成関数を使って標本集団を生成
3. 推定・検定・モデル選択などなど，統計的手法を適用し，標本集団から母集団に関する正しい情報が得られたか確認する

架空観測データ: 植物の花の数

「神の視点」で知ってること

- ポアソン分布からの無作為抽出
- 処理 B と無処理は同じ
- 処理 A だけが異なる

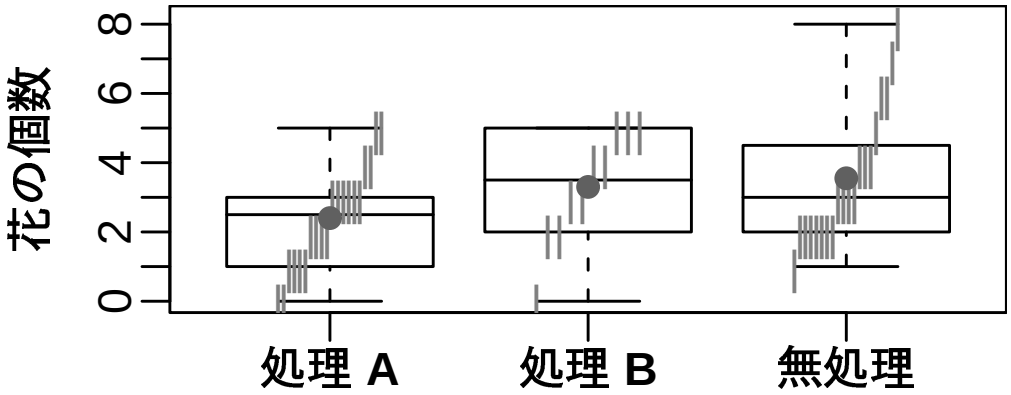
	記号	標本数	真の平均
処理 A	(A)	20 個体	2.5
処理 B	(B)	10 個体	3.5
無処理	(C)	20 個体	3.5



|, |: 一個体から得られたデータ, ●: 水準ごとの平均値

ここで考える問題と統計モデル

「人間の視点」で知ってること



- (A+B+C)?
- (A+B)(C)?
- (A+C)(B)?
- (A)(B+C)? (これが正解)
- (A)(B)(C)?

検討すべきこと

- 花の数の分布はどのような**統計モデル**で説明できるか?
- データにどうやって統計モデルをあてはめるか (**パラメーターの推定**)
- 処理 A ((A)) ・処理 B ((B)) ・無処理 ((C)) のどれが「同じ」でどれが「違う」と言えばよいのか? (「多重比較」? ここではすなわち**モデル選択**)

あなたのデータにぴったりの確率分布はコレ!

何でもかんでも変数変換しない

データにあわせて分布を選んで推定

— 選びかたの三つのポイント —

1. **説明したい量は離散か連続か?**
   - **離散**: { **生きてる**, **死んでる** }，**カウントデータ**, $\cdots$
   - **連続**: $\{0.56, 1.33, 12.4, 9.84, \cdots\}$，$\cdots$
2. **説明した量の範囲は?**
   - $\{0, 1, \cdots, N\}$, $\{0, 1, \cdots, \infty\}$, $[y_{\min}, y_{\max}]$, $[-\infty, \infty]$, $\cdots$
3. **説明したい量の分散 (ばらつき) と平均の関係は?**
   - **分散** $\approx$ **定数**, **分散** $\approx$ **平均**, **分散** $\propto$ **平均**, **分散** $\propto$ **平均** ^{n}, $\cdots$

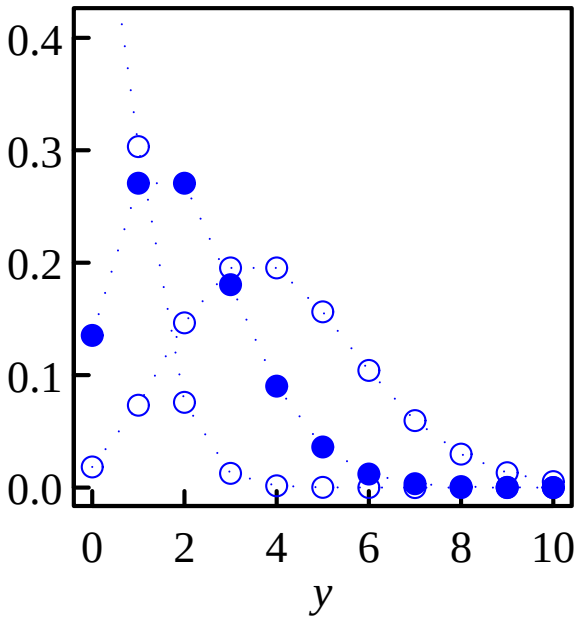
ポアソン分布 (Poisson distribution)

- **離散分布** $y_i \in \{0, 1, 2, \cdots, \infty\}$
- **確率密度関数** (paramter: λ)

$$\frac{\lambda^y \exp(-\lambda)}{y!}$$

- **期待値** λ**，分散** λ
- **使いどころ**:「一定時間にかかってくる電話の回数」……上限を設定できないカウントデータ
  - 産卵数・種子数
- **個数のデータが得られたら，まずは「ポアソン分布で説明できないか?」と考えてみる**

R の関数: `dpois`(y, λ)



確率分布を推定する方法たちの階層性

一般化線形モデルによって正規分布以外の確率分布を仮定した，パラメトリックな統計モデルたちを統一的に扱える

[**最尤推定法**で扱えるモデル]
何でもいいから確率分布があるモデル
一般化線形混合モデルなどなど

[**一般化線形モデル** (GLM)]
指数関数族の確率分布 + 線形モデル
ロジスティック回帰，ポアソン回帰などなど

[**最小二乗法**的に考えるモデル]
等分散正規分布 + 線形モデル
直線回帰，いわゆる「分散分析」などなど

一般化線形モデル (generalized linear model, glm( ))

- **指数関数族に属する確率分布あれこれ (正規分布，二項分布，ポアソン分布，$\cdots$) で説明されるばらつきのデータに適用できる**
- **link 関数を指定できる**
- **独立変数は何でもよい: 連続変数，名義変数，順序変数**
- **パラメーターは線形に結合していなくてはならない (線形モデル)**

$$\mathrm{link}(\mu(\mathbf{x})) = \beta_0 \cdot 1 + \beta_1 x_1 + \beta_2 x_2 + \cdots = \sum_i \beta_i x_i$$

統計ソフトウェア R では glm( ) 関数で簡単に推定計算，stepAIC( ) 関数で簡単に AIC によるモデル選択ができる

R と乱数と一般化線形モデル (glm())

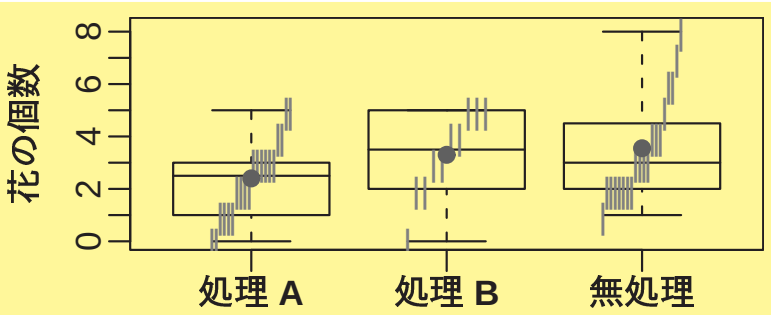
	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	rbinom()	glm(family = binom)
	二項分布	rbinom()	glm(family = binom)
	ポアソン分布	rpois()	glm(family = poisson)
	負の二項分布	rnbinom()	glm.nb()
(連続)	ガンマ分布	rgamma()	glm(family = gamma)
	正規分布	rnorm()	glm(family = gaussian)

- glm() で使える確率分布は上記以外もある
- glm.nb() は MASS library 中，またここには rnegbin() なども含まれる

1. 統計モデルを適用する標本集団を確定

(**データファイル**)

level	n.flower
A	3
A	0
A	1
…	…
B	4
B	1
B	2
…	…
C	5
C	4
C	2
…	…

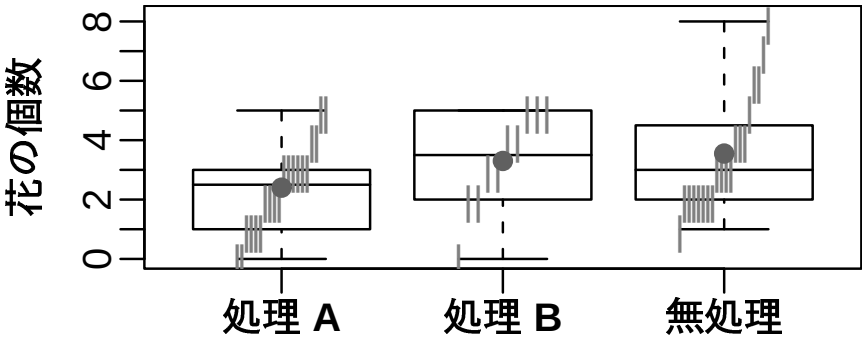


データ

- `level`: **水準**
  - A — **処理** A
  - B — **処理** B
  - C — **無処理**
- `n.flower`: **花の個数**
  - $\texttt{n.flower} \in \{0, 1, 2, \cdots\}$

2. 適用する統計モデルを列挙する

これは「三水準」 (A - B - C) **の問題である**



グループわけ	パラメーター数
(A+B+C)	1
(A+B)(C)	2
(A+C)(B)	2
(A)(B+C)	2
(A)(B)(C)	3

グループわけにともなう水準の「つけかえ」

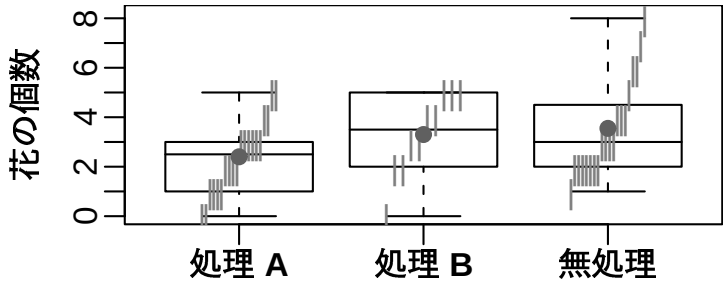
例: (A)(B+C) **ならば**

- A ⇒ (A)
- B ⇒ (B+C)
- C ⇒ (B+C)

level	n.flower	level.mapped
A	3	(A)
…	…	…
B	4	(B+C)
…	…	…
C	5	(B+C)
…	…	…
…	…	…

3. 統計モデルのパラメーターの最尤推定

一般化線形モデルのパラメーター最尤推定値を得るために，R の `glm()` 関数を使う．



- **グループわけ** (A+B)(C), (A+C)(B), (A)(B+C), (A)(B)(C) **の場合の** `glm()` **よびだし**:

```
glm(n.flower ~ level.mapped - 1,
    family = poisson(link = log))
```

- **グループわけ** (A+B+C) **の場合の** `glm()` **よびだし**:

```
glm(n.flower ~ 1, family = poisson(link = log))
```

4. モデル選択基準を計算する

これが小さいほど「良い」モデル

$$\mathrm{AIC} = -2(\textbf{最大化対数尤度}) + 2\,(\textbf{パラメーター数})$$

あてはまりぐあい (良い)　モデルの複雑さ (悪い)

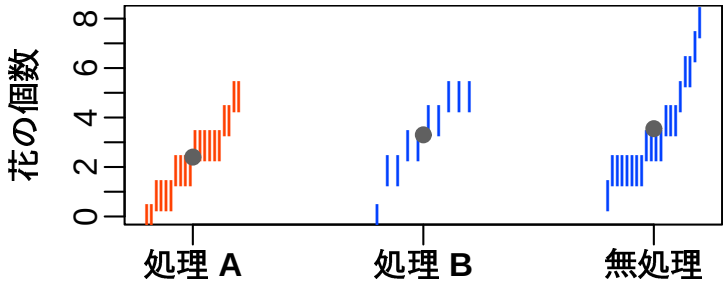
```
> result <- glm(n.flower ~ level.mapped - 1, family = poisson(link ...
> result  # 結果の表示
Call:  glm(formula = n.flower ~ level.mapped - 1, family = poisson(...

Coefficients:
  level.mapped(A)  level.mapped(B+C)
            0.875               1.243

Degrees of Freedom: 50 Total (i.e. Null);  48 Residual
Null Deviance:      189
Residual Deviance: 50.2         AIC: 193
> result$aic  # AIC の表示
[1] 192.91
```

5. モデル選択基準が最良のものを採用する

すべてのグループわけについて AIC を計算する．

```
> results <- estimate.poisson(samples)
> cat(sapply(results, function(r)
  sprintf("Model %-12s, AIC = %.1f", r$tag, r$glm$aic)),
  sep  = "\n")
```

```
Model (A+B+C)     , AIC = 195.5
Model (A+B)(C)    , AIC = 194.7
Model (A+C)(B)    , AIC = 197.3
Model (A)(B+C)    , AIC = 192.9
Model (A)(B)(C)   , AIC = 194.8
```



AIC **最小のモデルを選ぶ**

(注) ここで使ってる `estimate.poisson()` は久保の作った関数 (グループごとに `glm()` を呼びだし，最尤推定をおこない，AIC を計算する) で，R 標準搭載の関数ではない．

まとめ: 多重検定とどう違ったか?

モデル選択の利点: 簡単

familywise の危険率など計算しなくてよい

モデル選択の利点: 「矛盾」が生じない

多重検定では「$A = B$ の帰無仮説と $B = C$ の帰無仮説が棄却できない，しかし $A \neq C$ の有意差あった」という状況が生じうる; モデル選択ではこういった解釈の難しい結果はでない

モデル選択の欠点: 過誤の確率を統制できない

「やくざ」むけ —- つまり基礎科学研究むけ

Rで強める多重比較なモデル選択

Rプログラミングで難しい問題も解決できる

もし処理の水準がもっと多かったら?

めんどうなことは R にやらせる— 計算機に使われるのではなく計算機を使う

generate.groups() **関数による「すべての可能な組み合わせ」の生成** (4 **水準**)**.**

```
> sapply(generate.groups(c("A", "B", "C", "D")), function(g) g$tag)
 [1] "(A+B+C+D)"    "(A+B+C)(D)"   "(A+B+D)(C)"   "(A+C+D)(B)"   "(A)(B+C+D)"
 [6] "(A+B)(C+D)"   "(A+C)(B+D)"   "(A+D)(B+C)"   "(A+B)(C)(D)"  "(A+C)(B)(D)"
[11] "(A+D)(B)(C)"  "(A)(B+C)(D)"  "(A)(B+D)(C)"  "(A)(B)(C+D)"  "(A)(B)(C)(D)"
```

今回の自由集会のために作った便利な (?) 関数たち

- partition.int(): **整数の分割**
- generate.groups(): **グループわけの列挙，水準の射影**
- estimate.poisson(): generate.groups() **よびだしつつ，一般化線形モデルによる推定を行なう**

本日のまとめ

1. **「検定」すべきかどうか，よく考えよう**
   - 第 I 種の過誤だけが重要か? ホントに「帰無仮説」?
2. **モデル選択で簡単に** — 検定に比べて理解しやすく解釈しやすい
   - 「多重比較」 → グループ化，最尤推定，AIC 計算
3. **R でプログラミングしよう** — あるいは誰かに作らせる
   - 必要なものは自分で作る，解析手法をデータにあわせる

本日は説明しなかったこと

- 水準間に順番ある場合 (可能な場合わけの個数が減る)
- 「個体差」がある場合 → 混合モデル，など
- もちろん R でも「かたぎ」な多重検定はできます (`multcomp` library など)

2004.08.28