データ解析のための統計モデリング (2008 年 10-11 月)

全 5 (+2) 回中の第 1 回 (2008–10–27)

生態学データ解析の統計モデリングとは?

久保拓弥 kubo@ees.hokudai.ac.jp

http://hosho.ees.hokudai.ac.jp/~kubo/ce/EesLecture2008.html

この講義のーとが「データ解析のための統計モデリング入門」として出版されました!

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

まちがいを修正し，詳しい解説・新しい内容をたくさん追加したものです

今日のもくじ

この統計学講義のーと [1] は

- 私 (久保) が講義でしゃべることをど忘れするのを防ぐ
- 皆さんがノートをとる手間をはぶく [2]
- 皆さんがあとで読みかえして復習する
- ネット上で誰でもダウンロードできるようにする [3]

ために準備しました．まあ，この講義のーともいろいろと不備なんかがあると思いますが，ネット上のこの講義ページ (URL は目次の上に，入力がめんどくさければ「久保 講義のーと 2008」で検索してください) でまちがい修正だとか随時更新していければ，と考えています．また，次回以降の講義ノートもこのへんにアップロードするので，それをあらかじめダウンロード

1. 教科書というよりもっといいかげんな講義のーとです.

2. 左右の余白は注釈・メモ・らくがきのためのものです．手を動かしてないと眠くなりますよ.

3. しかしこの教室にいないとよくわからない内容かもしれませんねえ.

& 印刷出力してもらうと，当方としても助かります．これはこの講義専用のメイリングリストで連絡します．

この「生態学の統計モデリング」って講義でいったいぜんたい何をやるつもりなのか? そうですねえ，スローガン的には

- これからデータ解析をする大学院生たちのために
- 実際のデータ解析で使う**統計モデルの考えかた**を
- 基礎から最先端につながるような**「地図」**を示しつつ
- **理解できる**統計学めざして

といったモノになってくれればなぁ，と．こうやって皆さんの前でしゃべってる私のほうはそういうつもりなんだけど……実際のところ，聴いてる皆さんのほうはなかなかすんなりとは理解できないかもしれません．ということで，わからないことがあったらどんどん質問してください． [4]

4. しかし院生の皆さんに質問を発してもらうのは，ときとしてかなりの難事だよなぁ，という気分だったり．

1. 環境科学にとって統計学的なデータ解析とは?

この統計学モデリングの講義では，おもに生物の環境科学 [5] というか生態学 (ecology) なんかの研究をしている大学院生に説明しているつもりでハナシをススめます．生態学固有の知識なんかはできるだけ持ちださないでハナシをしたい，と考えています．しかしながら，ときとして「個体群」とか生態学用語なんぞを使ってしまうかもしれないので，これまたわからならかったら臆せずに質問してください．

5. これはいちおう「環境起学」専攻とやらの講義なんで……

さて，ハナシの順番からいって，まずは**「そもそも何で統計学だの統計モデリングだのが必要なのか?」**といったところから始めてみましょうか．

生態学では野外・室内での観測や実験によって，自然現象から何か情報を取りだそうとするんですけど，このときに二段階の**情報消失・情報圧縮**が発生します (図 1)．つまり……

- 第一段階: 自然現象から観測・実験といった手段で情報をとりだして「データ」(数値・記号の集まり) に変えるときに多くの情報が失われる
- 第二段階: データを統計解析するときに多くの情報が失われる

まず第一段階では，生きている生物がもつ莫大な情報を人間が観測・測定してみることはほぼ**不可能**，それを実験室の簡単な条件下においても，ということです．たとえばそこらへんの地面に生えてる草について考えてみましょ

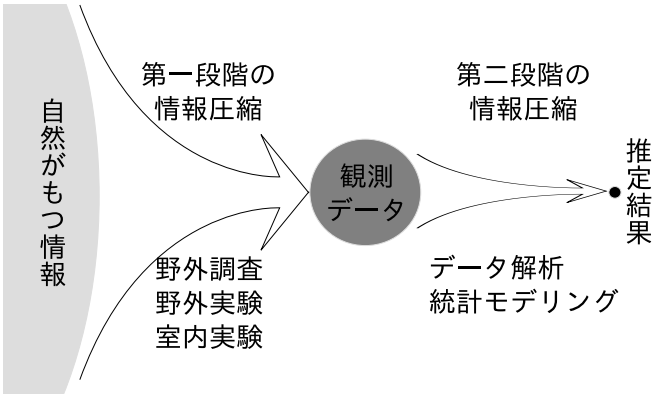

図 1: 生態学における二段階の情報消失・情報圧縮

う．高さを測定したり葉の枚数を数えたり光合成速度を計ったり，といったことは可能でしょうし，刈り取ってしまえば重量を測定したり，遺伝子を測定したりといったこともできるでしょう．

しかし何もかも調べることはできません．野外観測データをあつかったことがある研究者なら，われわれ人間はさまざまな観測記録機器を駆使しても自然がもつ情報の**ごく一部しか「データ化」できない**，ということに同意してくれるだろうと思います．

何やらくだくだしく説明してきましたが，この講義ではこの「第一段階の情報消失」については検討しません．そこまでいくと各分野での観測の方法論だの「研究によって明らかにしたいこと」だのにまで踏みこまねばならないでしょうし，そんな面倒は私も皆さんも回避したいところでしょう．

この講義では第二段階の情報消失，つまり**統計解析によって情報が失われる**，良くいえば「圧縮」される問題に集中したい，と考えています．[6] 自然現象に含まれていたはずの莫大な情報のごくごく一部が数値と記号の羅列すなわち「観測データ」とやらにいったん変換されてしまったならば，どんな学問分野で得られたデータであろうと，たいていの場合は統計学的な手法を使ってそれを調べるほかないからです．これはじつに驚くべきことなのかもしれませんね．

6. ホントはこの統計解析の問題は第一段階にも影響する，つまりどう自然を観測・測定するか，にも密接に関わるのですが，ハナシが広がりすぎるのでやはり省略します．

で，豊潤な自然現象の「しぼりかす」みたいな観測データを何でまたわざわざ統計学的手法とかで情報消失・圧縮してやらねばならないのか，という問題なんですが……ヒトことで言えば，われわれ**人間はアタマが悪すぎて**データなる数値・記号の羅列をうまく把握できないからなんです．私なんかだとデータ量が 3 行 × 3 列ぐらいをこえると，もう何がなんだかよくわからなくなったりします．

ということで,「おそらく自然現象の一部を反映してくれているであろう[7] 観測データ，その観測データに見られる[8] さまざまな『パターン』をうま

7. これは単なる願望であることが多いんですけどね

8. あるいは見えてるのかどうかもよくわからない

く説明できるような数理モデルがほしい」という状況になります．数理モデルってのは数式で書かれた比較的簡単なモデル，さらにそのモデルに「測定誤差」だの「個体差」だの「不確定性」だの「変異性」だのが混入してしまっているもの，つまり確率論的な数理モデルのことをこの講義では**統計モデル**とよぶことにしましょう．私が言いたいことをまとめてみると，

- 「自然科学という方法で自然現象を理解する」なる営為は結局のところ「情報を捨てることで人間でも理解できる [9] ようにする」こと
- とくにいったん「データ化」してしまえば，そこからの「情報捨て」は統計学的な方法であつかうしかない— そしてここから先の情報処理は“客観的”，つまりあなたがやったことを私もまた同じように**再現**してみせることが簡単である，ということです
- とくに (実験による現象の再現が難しい) 生態学など環境科学の研究者の世界では「自然を観察・観測して言えること」の少なくない部分は統計学的方法に依存していて， [10] 研究の信頼度を上げるための研究者の相互検証はこのデータ解析の再検証によることが多い

後述するように計算機ハードウェア・ソフトウェアの進歩によって，1990年代から統計学的な道具は大きく発展しました．生態学まわりでのデータ解析の方法も大きく変わってきました．いわば「りすとらされたデータ解析の世界」で，何から勉強していけばいいのかをこの講義で考えていきましょう． [11]

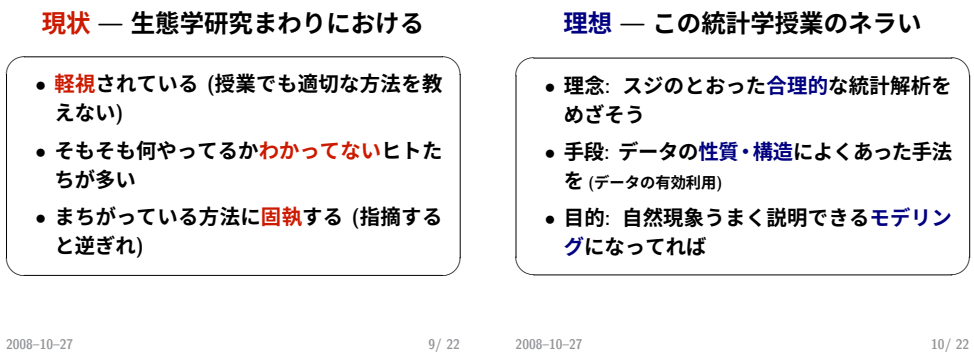

図 2: これまでの状況，この講義でめざすところ

9. もしくは理解した「気分」になれる

10. もちろん現場で自然現象をみて「自分はここでこういうことが生じていると思った」という直接観察は重要です……しかしながら，その思索だけでは自然科学には含まれないんですね．自然科学では数値化・記号化されたデータこそが (共有できるという意味で)「客観的なものだ」という共同幻想を維持していますので，こいつを「客観的」にモデリングしてパターンを抽出する必要があるのです．

11. このへんの導入部分では「ヤミにほうむられる修士論文」とデータ解析の関係，といったことも検討したかったのですが，時間ぎれで書けませんでした．これはまた別の機会に……

2. 統計モデルって何?

「観測データにもとづいて『何か』を主張する」ためには統計学的なデータ解析がたいへん重要な役割を果たすはず— にもかかわらず，生態学まわりでの統計学の使われかたを見ていると，ちょっと前までは何ともおそまつなものでした (図 2 の左, あるいは Appendix でぐだぐだと書いていること).

現状は必ずしもよろこばしいものではない，それがゆえに，この講義では「理念: スジのとおった」「手段: データをよくみて」「目的: 現象を説明する」統計モデルをめざしましょう (図 2 の右)，といったあたりを**身につけてもらいたいデータ解析の姿勢**としてかかげてみたいと思います．なんだかおかたいハナシのようですが，観測データをこうやって解析していくのは**じつに楽しいこと**という点も強調しておきます．[12]

さて，ここで重要になるのは**統計モデル**いった考えかたです (図 3)．統計モデルってのは何なのかを説明してみるとこんなかんじになるかと思います．

- 観測データを解析する，ってことは**統計モデリング**である
- 統計モデルは観測データのパターンをうまく**説明**できるようなモデル
- 統計モデルの基本的な部品は**確率分布**，確率分布のカタチはパラメーターによって決まる
- 観測データをうまく説明できるようにパラメーターの値を決めることを「統計モデルの**あてはめ**」または「統計モデルによる**推定**」という[13]

各項がのべようとしていることは，この講義の中でおいおい説明してみるつもりです.

12. これに対して，スジのとおらない，何をやってるのかワケのわからない，現象を説明できてるのか判然とせぬデータ解析はまったくカナしいものです.

13. 観測データとの対応関係が明確ではない「数理モデル」をあつかっているヒトたちもいます．こういったモデルは，(1) 観測データからパラメーターを推定する手段，(2) モデルの予測とデータの乖離を定量化する手段，のあいだに整合性がありません．つまり観測データにあらわれるパターンを説明したいのであれば，(1) と (2) を最初から考えている統計モデルのほうがずっと有用なのです.

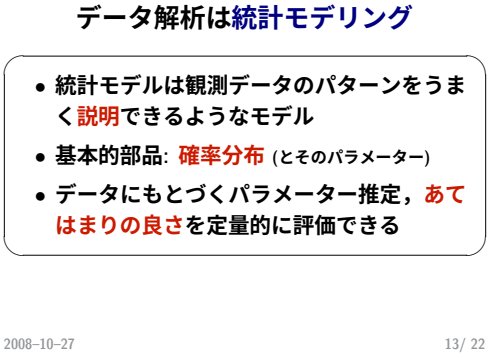

図 3: 統計モデルとは何か?

最初に一点だけ強調しておくなら，誰もがデータ解析をやるときに統計モデリングを回避できない，というあたりでしょうか.「もでるぅ？ オレは自然現象そのものを調べてるから数理モデルなんかとは何の関係ないもんね」てなことを言ってる**カンちがい**ぎみの研究者であっても,「まあこの論文にのせるデータ，そのまま table にしても reviewer に文句言われるから『いちおう検定にかけて』[14] 『ゆーい差』でも出しておくか」などと行動するのであれば……このヒトは，確率分布をくみあわせた統計モデルを観測データにあてはめてみて，パラメーターだの検定統計量だのを推定している，ということになるのです．つまり自分が何やってるのかわかっていない，そういうヒトたちが統計学的手法ユーザーのたいはんをしめています.

この講義の目的としては，皆さんがいかなる状況にあっても「自分はいま**どういう統計モデルを**どうあつかっているのか，観測データとはどう対応しているのか」を明確に意識しつつデータ解析にとりくめるようになる，てなことがこの講義のゴールで実現できればよいのだろう，と考えています.

14. これはじつにひんぱんに聞かされる意味不明な表現……というか「ゆーい差」にかかる枕詞?

3. データ解析の勉強のやりかた

近ごろの北大の修論発表会や生態学会の大会の研究発表をみていると，**大学院生たちのほうが統計学をよくわかっている**，その指導教員たちにくらべて，といった気分にさせられます．そうなる理由としては，大学院生のほうが統計学をよく勉強していて，しかも図 4 に示しているように現在は統計学の勉強をススめていくのに良い条件がそろっているためではないでしょうか.

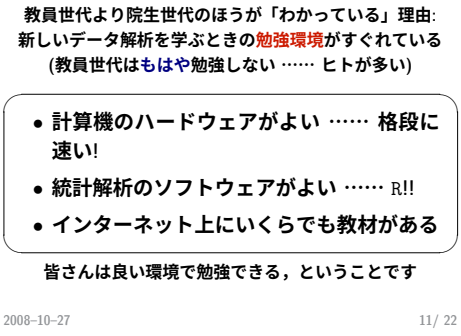

図 4: 今どきのめぐまれた勉強環境

簡単に説明してみますと……そもそもデータ解析は計算機のパワーが必要であり，この授業であつかうような手法は現代の「パソコン」ぐらいの計算能力がどうしても必要となります.[15] また計算機のパワーを使いこなすためには良いソフトウェアが必要であり，(後述する) R のような優れた統計ソフトウェアが無料で簡単に入手できる，というのも昔なら信じられないことです．さらにインターネットを介して「世界中の統計学講義のーと」が読めるようになりました．[16] 数式ばかりの統計学教科書にひたすらとりくむ他なかった教員世代とくらべると，現在はずいぶんと勉強しやすい環境になったといえます．とくにデータを使っ

15. これは 20 年前であれば数億円の大形計算機に匹敵するものです.

16. これまでの久保「講義のーと」のたぐいも数百人 (もっとたくさん?) のヒトにダウンロードしていただきました.

て試行錯誤，つまり手をうごかしてデータの図を作ったりしながら概念を理解していくことはとても有用です．ですから，皆さんもこの勉強環境をうまく活用してみてください．

この講義では「ふつーの統計学の講義」では扱わない「基本から最先端までを橋渡しする統計モデリングの流れ」を重視します．一方で，ふつーの講義で扱われることを，あまり説明しない傾向があるかもしれません (図 5)．そこでこの「欠けた部分」をおぎなうために，皆さんに統計学の本を買ってもらってそのあたりを自分である程度は勉強する必要があります．

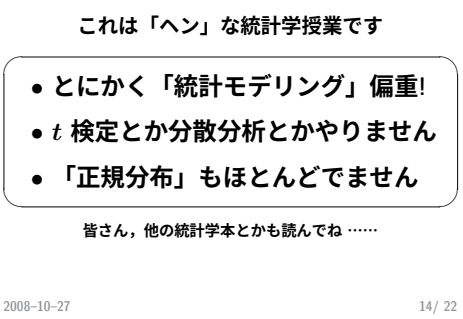

図 5: 「ヘン」な講義にようこそ

さきほど述べたようにインターネット上でいろいろな統計学の講義資料などが配布されています．この講義の欠落部分をおぎなうためにはそういったものを読むことも有用だろうと思います．

また世の中には多数のすぐれた統計学本がでています．これらもぜひ活用してください．図書館や本屋でいろいろとぱらぱらと見てもらって，自分にとってわかりやすそうな入門的な教科書が一冊てもとにあればずいぶんと勉強の役にたつだろうと思います．とくに図の多い統計学教科書は重要で，その意味では (もしわかりやすければ) 「マンガでわかる……」のたぐいの統計学本も良いかと思います．

ここでは二冊ほど紹介してみます．まず一冊目は**「統計のはなし」** (粕谷英一, 1998, 文一総合出版)，生態学研究者のあいだでは「ぴんく本」として知られるものです．これは読みやすく，また生態学まわりでの統計学的手法が概観できて，とてもおすすめの一冊です．巻末付録「君にもできるごまかし」もいろいろな意味でたいへん勉強になります．[17]

17. このぴんく本に影響されたヒトはたくさんいるので，皆さんの指導教員のアタマの中がどうなってるのか，それを予測するときにもたいへん有用です．

もう一冊は**「統計 (第二版)」** (竹村彰通, 2007, 共立出版) です．統計学の専門家が書いた「統計的手法の使いかたを示しながら理論を説明」するもので,「統計学を学ぶ際の動機づけ」を重視して「統計学の手法を用いるとどんなことができるか」を示しています．またここでは統計ソフトウェア R による解析例が示されています．

データ解析の手法の勉強するときに実践したら良いことを図 6 にまとめてみました．私がおすすめしたいのは実際にデータをあつかいながら，とくに

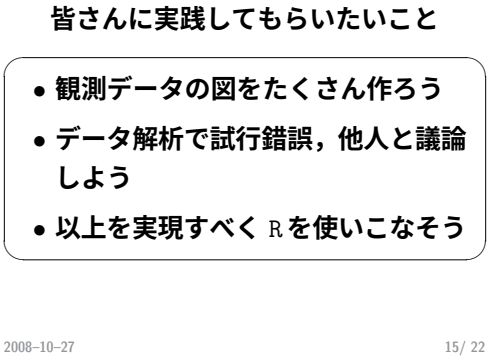

図 6: データ解析の勉強するときに…

図を描くという作業をしながら勉強をススめていくことです．そして他人との議論も重要です．それは人によって「データを使っての試行錯誤」体験がずいぶんと異なっているからで，そのあたりの自分・他人の「ずれ」から多くのことが学べます．[18] ひきこもっていてはデータ解析の勉強はできません．このように勉強をススめるとき，そしてもちろん研究のデータ解析をススめるときに有用なのが R と呼ばれるソフトウェアです．

18. 私はいつも大学院生たちからの質問，とくにふうがわりな質問に勉強させられます…… いや，ホントにそう思います．

4. データ解析ソフトウェア R って何?

この授業では統計ソフトウェア R の利用を前提にハナシをすすめていきます．[19] みなさんも自分の PC にさっそく R をインストールしてください．インストール方法は

`http://hosho.ees.hokudai.ac.jp/~kubo/ce/HowtoInstallR.html`

なんかが参考になるだろうと思います．[20]

R は統計ソフトウェアは実際の解析だけでなく，統計学的な手法を勉強するときにも必要なものです．[21]

いろいろあるソフトウェアの中で R を皆さんにおすすめします．というか，この講義の中では R を使うことを前提にハナシをすすめます．R の利点は

- **free software** である — つまり誰でも無料で入手でき，しかもそのしくみ (source code) が完全に公開されている [22]
- 強力無比な作図機能がある — **「データを図にする」**ことはデータ解析においてとてもとても重要 [23]
- プログラミング可能

19. R については，今回 (印刷出版前にこっそり) 配布する「R で改善する生態学のデータ解析」も参照してください．

20. このあたりの URL などはあとでメイリングリストでおしらせします．

21. いろいろな方法を適用する「実験」ができますからね．

22. このことは学術研究においては特に重要です．

23. この講義の中でもそれを強調していきたいと思います．

- add-on package library が充実していて，機能拡張が容易

といったことです．データ解析の勉強のおともとして，あるいは自分の研究の強力な仲間としてたいへん頼りになるソフトウェアです．

5. この全 5 回の授業では何をあつかうのか?

この講義では R を駆使して,「自分のアタマで理解して使う統計学」をめざしたい，と考えています．同時に，観測データを不自然にひねくりまわさない,「素材のよさ」をいかしたデータ解析ができるようになってくれれば，とも考えています．

そのあたりを実現するために，今回の全 5 回の統計学の講義は

- 確率分布で表現される統計モデルを重視する
- 統計モデルがデータにどれだけあてはまっているかを尤度(ゆうど) で評価する
- 線形モデルを中心にあつかい，これを一般化線形モデル → 一般化線形混合モデルと拡張していく [24]

という基本方針で，図 7 に示しているような統計モデルの「地図」の内側から外側にむかってススめていく予定です．この図をみれば何となくわか

24. といった予定だったのですが，後述しているように一般化線形混合モデルの部分は消えました．

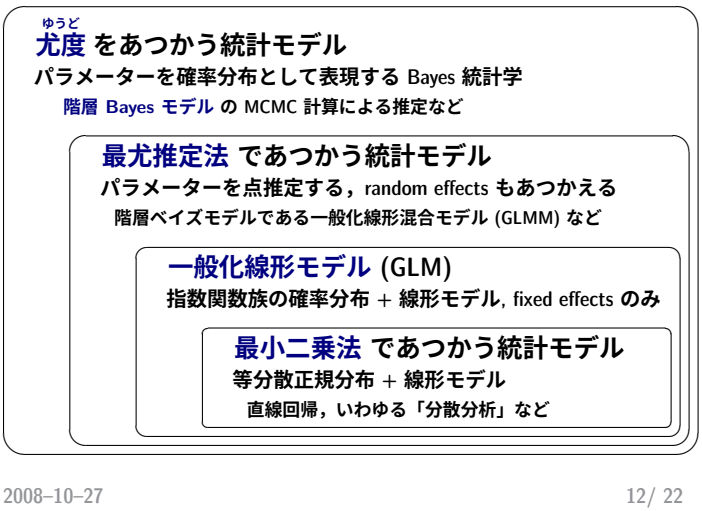

図 7: 統計モデルの「地図」

るように，上でちょっとマンガふうに描写した「指導教員たちの統計学の世

界」ってのは図の中のごく狭い領域「最小二乗法であつかう統計モデル」に限定されている，といったことを示しているつもりです．この講義では，そういう「最小二乗法であつかう」狭い領域は**一般化線形モデル** (generalized linear model; GLM) の一部にすぎない，というあたりも説明できれば，と考えています．

とりあえず全 7 回の流れに対応させてこのあたりの概要を説明してみましょう．

1. 10/27 (月) 生態学データ解析の統計モデリングとは?

今回ですね．統計モデリングの必要性，生態学における今までの統計学的手法の使われかたについて説明してみました．このあと，統計学勉強に必要な本とソフトウェアについて説明するつもりです．

2. 10/30 (木) さまざまな確率分布と最尤推定

次回は統計モデルの「基本部品」というべき**確率分布**について説明します．世の中にはたくさんの確率分布があって，ホントはいろいろ紹介したいのですが，この講義で使う 3 つの確率分布，すなわち「ポアソン分布 (Poisson distribution)」「二項分布 (binomial distribution)」「正規分布 (normal distribution / Gaussian distribution)」についてだけ説明する予定です．またデータへの「あてはまり」をあらわす**尤度**(ゆうど)について説明し，これまでの統計学のパラメーター推定の基本的な考えかた，最尤(さいゆう)推定法についても説明します．[25]

25. 11/3 (月) は祝日なので講義はありません．

3. 11/06 (木) 一般化線形モデル (GLM) 1

ポアソン分布で説明されるであろう現象[26]の統計モデリングについて説明します．ポアソン分布モデルのパラメーターを最尤推定することを，ポアソン回帰といったりします．一般化線形モデル (GLM) というのは (線形) ポアソン回帰とかこの次の回の (線形) ロジスティック回帰など線形モデルのたぐいをいっしょくたにして呼ぶときの総称です．

26. 観測データが「0 個, 1 個, 2 個，...」というふうに整数として数えられるカウントデータであるとき，など．

4. 11/10 (月) 一般化線形モデル (GLM) 2

よく使われるもうひとつの GLM である**ロジスティック回帰**について説明します．これは「生物が刺激に応答した・しなかった」とか「観察された N 個体のうち y 匹が死亡した」といった「あり・なし」データを解析する統計モデルにもとづいていて，この統計モデルは**二項分布**を基本部品とする統計モデルです．

5. 11/13 (木) 検定とモデル選択

ある観測データを説明できる統計モデルはたいてい複数あります．このときに複数の統計モデルをどう関係づけるのか，伝統的な Neyman-Pearson わくぐみの「検定」の正体を説明します．さらにまた別のモデル評価法，情報量規準をつかったモデル選択についても説明します．

6. (消された講義その 1) 一般化線形混合モデル (GLMM)

データ解析で GLM を使いこなせるようになってくると，GLM ではうまく説明できない現象あるいは観測データ内のパターンがあることに気づきます．これらは「個体差」「場所差」など人間が測定してない / 測定できない要因によって生じるばらつきだったりします．そこで「個体差」「場所差」を **random effects** として GLM に組みこんで一般化線形混合モデル (generalized linear mixed model; GLMM) に拡張します．

7. (消された講義その 2) 階層ベイズモデル

データ解析に GLMM が使えるようになると，今度は観測データのあちこちに random effects になりうる要因があることに気づきます．たとえば野外観測データの場合,「場所差」があってその中に「個体差」がある，といった状況です．このように統計モデルが複雑になってくると，最尤推定法によるパラメーター推定がだんだん「苦しく」なります．そこで，GLMM を階層ベイズモデルの一種だとみなして，**Markov Chain Monte Carlo (MCMC) 法**を使ってパラメーターの事後分布をもとめる方法を紹介します．

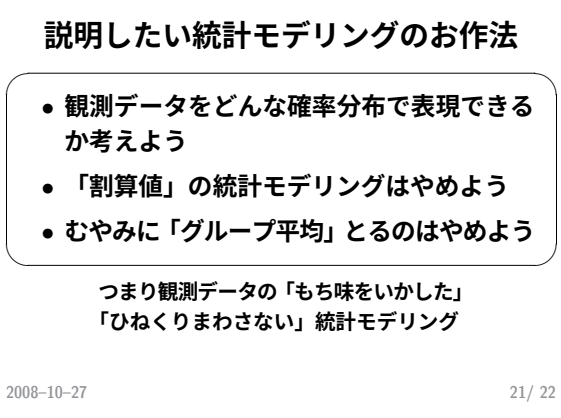

図 8: この授業で強調する統計モデリングの「お作法」

最後にもう一度この授業のネライみたいなものをまとめてみましょう．私がおススめしてみたい**データ解析の姿勢** (図 2 の右) としては,「理念: スジ

のとおった」「手段: データをよくみて」「目的: 現象を説明する」**統計モデリング**をめざしましょうといったあたりです．**データ解析の勉強** (図 6) としておススめしたいのはデータを使った試行錯誤と他人との議論です．そして**統計モデリングお作法**としては，図 8 にまとめているように,「確率分布の重視」「なんでも割算するな」「なんでもグループ化するな」といったところです．次回からはこれら三つの注意点を実現できるようにハナシをすすめていきます．

とりあえず，今日はここまでにしましょう．

A1. これまでの生態学における統計学の使われかた

本文で「もっと観測データと対応のつく数理モデルにしろ」とか言ってたヒトたち自身がやっているデータ解析は，けっこう悲惨なモノとのべました．実際のところどういうものなのか,「何をやってるか」といういうより「どのようにやってるのか」あたりの**行動観察**の結果を示してみましょう．以下の流れは，かなり多くの生態学研究者で見られるものです． [27]

1. 自分と同じようなデータをとってる論文をひたすらさがす
2. 見つけたら「どういう『検定』を使っているか」をチェックする (これを検定 X としましょう)
3. 「統計ソフトウェアで検定 X の結果として○○という output が得られた → その場合は△△と断定してよい」というパターンをひたすらまる暗記する [28]
4. 検定 X ができる統計ソフトウェアを買ってくる
5. 統計ソフトウェアにデータを流しこむ
6. 検定 X つかった output が得られたら 3. で暗記したパターンを適用して論文をかき，投稿する
7. reiviewer が「検定に文句をつけたら」，reviewer が指定する検定を新しく検定 (新) X と指定しなおして，ひたすら 3.- 7. をくりかえす [29]

何とも「悪しき行動ぱたーん」と言いたいところですが， [30] ここまでくるとむしろ何かこっけいなもののように思えてしまいます．もちろんやってる当人はごく真摯なもので，むしろ上のような行動パターンこそが「データ解析の必勝パターン」とでも信じているのでしょう．

上の一連の行動では何やら「検定」 & 「ゆーい差」なるもの [31] ばかりが重視されているみたいですが，この部分を極限まで押しススめてしまうヒトたち [32] なんかもいたりしまして，そういうヒトたちのが示している奇妙な行動をちょっと列挙してみましょう．これらもまたことごとく私が実際に目撃・観察した事例の一部です．

- 「ゆーい差」が出るまで「検定」手法をひたすらとりかえる
- データ中の観測値どうしを割算したりして新しい「指標」をでっちあげる…… 「ゆーい差」がでるまで新発明・珍発明をくりかえす
- 都合の悪いデータをとりのぞく
- 「ゆーい差」が出るまでデータサンプリングをひたすら続ける

27. そしてこの教室にいる大学院生の多くも，修論とか書くときにはこういう行動とるんじゃないか，という悲観もあります．

28. そして，これらはだいたいにおいてまちがった「チャート式」になってたりするんですよね．

29. reviewer が指定する検定 (新) X をやるためだけに，十数万円もする統計ソフトウェアを新しく買う……私が目撃した実例のひとつです．

30. あるいは「検定 X の正体が何なのか，まったく理解してなくても論文が書けてしまう」ひどく巧妙なる方法論，と言えるのかもしれません．

31. これらはあとの回の講義でとりあげます

32. 「ゆーい差」決戦主義者…… まあ，これにもいろいろなタイプがあるのですが．

このヒトたちがここまでして「ゆーい差」にこだわるのは

- 「ゆーい差」さえ「出せれば」，あとは何でもかんでも好き勝手に断定してみせてよい，という思いこみがある— 逆に「ゆーい」じゃないと不安で不安でしょうがない
- 統計学のこととかよく知らないけど,「ゆーい差」ならわかる [33]

といったキモチがあるようです.

33. 実際には「ゆーい差」についてもたいしてわかってないことが多いです.

また，大学院生にとって多くの場合，自分の研究の方向性を左右したり，データをどうあつかうかを決めるのは指導教員ですよね……ということで，その「ありがちな統計学的世界」を知っておくのは有用だろうと思います.

以下に示されてる図式は**ほとんどまちがい**なので，その意味するところがわからなければマネしないほうが安全だろうと思います.

- 「検定」はたいていの場合，正規分布を仮定すればよい
  - 値 X vs 値 Y みたいなデータはことごとく直線回帰すればよい
  - 「処理区」「無処理区」というデータは ANOVA すればよい
  - しかしこれらはどう違うのか理解できない
  - それはともかく，こういう解析やったときに統計ソフトウェアが吐きだす R^2 値は「説明力」で，それが大きけりゃなんでもいい
- 「等分散じゃない」とか文句をつけられたら，データを対数変換だの arcsin 変換して回帰や ANOVA すればよい
  - あるいはめんどうになったら割算値とかぽんぽん作って「のんぱらめとりっく検定」やればよい
- 「検定を何度もやってる多重比較だ」と文句つけられれば，何でもかんでも多重検定法による p 値補正をやればよい
- データがちょっと複雑な table であれば，何でもかんでも χ^2 検定やればよい
- もっとめんどうなデータは何でもかんでも多変量解析やればよい

実際のところ，こういうヒトたちのたいへんなところは，上にあげたことの当否というより,

- 上のようにデータ解析して論文が accept されてきた，つまり自分自身の**「成功体験」だと信じている**

- とにかく上のような「世界」を固守していればよい
- それゆえに，もはや統計学の勉強なんかする必要がない

つまり「アタマがかたい」というべき状態にあることかもしれません.

大学院生にとって重要なことのひとつは「大学院という牢獄」から脱出することなので，指導教員には逆らわないでその言いなりになって研究をまとめて卒業する，というのもかなり現実的な方策だろうとは思います．そういう大学院生たちには，残念ながらこの講義はそれほど有用ではないかもしれません……というか統計学の勉強なんかする必要ありませんよね.

A2. 私はどのようにデータ解析の専門家になったか

この節ではどう勉強したかというより，まあ，私の研究者としての背景，みたいなところを説明してみましょうかね.

私はデータ解析つまり統計モデリングをもっぱらとしていて，観測データにみられるいろいろなパターンをうまく説明する，という部分が現在の専門といえるでしょう.

しかしながら，もともとそういう統計学なんかの教育をうけたわけではなくて，大学院生として所属していたのは「数理生物学」なるものを専門とする研究室でした．この数理生物学ってのは現実ばなれというか，観測データとの定量的な対応とかぜんぜん気にしてない分野です．[34] もちろん数理モデルを使えばいろいろなパターンを生成できますから,「やあ，何だかこれは自然現象に見られるパターンと『似てる』ぞ」とか学会発表してみたりしたんですけど，実際にホンモノの生物・生物集団をあいてにしてるヒトたちからは「ふーん，へー」とあまり感心されない．親切なヒトなんかはわざわざ「もっと観測データと対応のつく数理モデルにしろ」とか言ってくれたりしました.

そりゃ，たしかに当方が「似てる，似てる」などと騒いでみせたところで，それはぜんぜん定量的なハナシでも何でもありません．そして，観測データなるものを定量的にあつかおうとすると……これはやはり統計モデリングするしかない，ということに大学院生だった私でもいやおうなく気づかされたわけです．で，独学で統計学の勉強なんぞやってたりすると，数理生物な研究室のヒトたちからは**「データ解析なんかやって何がわかるんだ!」**とか言われたりするんですよね．[35]

ともかく，統計学を勉強して野外なんかの観測データをあつかった論文をすみずみまできちんと理解して読める[36] ようになってくると，ですねえ……

34. なんというか,「身内」だけに分かるいろいろな方程式のたぐいをヒネくりまわすことばかりやっている分野，と戯画化できるのかもしれません.

35. 実話です.

36. ありていに言ってデータ解析の手法が理解できていない人は自然現象をあつかった論文をごく不十分にしか読めていません.

「もっと観測データと対応のつく数理モデルにしろ」とか言ってたヒトたち自身がやっているデータ解析ってのは，じつはかなり**悲惨**なモノだといったことが理解できるようになりました．このヒトたちはなるほどデータを得るのに多大な苦労したんだろうけど，それを解析の段階で「めちゃくちゃ」にしてしまっている，と．[37]

37. じつは他ならぬ私自身なんかも，昔の自分の論文とかみると「これはいくらなんでもヒドすぎる」と思うことが多々あるのですが……

このあたり，つまり生態学研究者たちの「統計学の使いかた」といった行動観察についてはまたあとで述べますが……個人的には大学院生ぐらいのころは「データ解析なんかやって何がわかるんだ!」と決めつけられる理不尽とか「もっと観測データと対応のつく数理モデルにしろ」と言いつつだめだめなデータ解析ばかり横行してる現状，に対する何というか反抗的なココロがデータ解析・統計モデリングの勉強をススめていく動機の一部でした．

しかし統計モデリングに関する理解が深まり，野外調査や実験をもっぱらとする研究者たちとの共同研究ができるようになって，その観測データからいろいろなパターンが抽出できたり，それをうまく統計モデルで簡単に説明できるようになってくると……まあ，上のようなヒトたちは取るにたらない，わざわざ相手するほどでもないヒトたちだとわかってきました．

そんなことよりも，データをとっているヒトたちとこういう統計モデリングの考えかたを共有して**いっしょに改善していく**，ということのほうがよほど重要だろう，と．とくに大学院生なんかはまだアタマが柔軟なヒトが多いですから，[38] いろいろ議論したりやりとりしてるうちにその指導教員なんぞは**あっというまに抜きさる**ほど統計モデリングに対する理解を深めたりしてくれる．これは何度みていてもホントにおもしろいことです……ということで，私としてはこの講義で何人かの大学院生やひとつデータ解析を勉強してやろうと思ってるヒトたちが，統計モデリングに関して今までよりすっきりとした理解ができるようになってくれれば，私にとっても皆さんにとってもおもしろいことだろう……といったコトなんかもこの講義をやる動機になっています．

38. 実際のところ，私のところにデータ解析質問メイルなんかをくれるのは，たいてい大学院生だったりします．