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 5 回 (2008–11–13)

検定とモデル選択

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

「生態学の統計モデリング」第 5 回目 (いちおう今回の最終回) です．さてさて，ここまでのハナシをふりかえってみると，

- データ解析は**統計モデリング**であり，統計モデルの基本的な部品は**確率分布**である
- 観察されたデータにみられるような統計モデルを作り，**最尤推定法**によってパラメーター推定する
- 統計モデルの中で**一般化線形モデル** (GLM) というのがなかなか便利に使えそうだ……とくにカウントデータの解析において

- GLM は (データの種類に対応する) 確率分布と **link 関数**をえらび, 要因をくみあわせて**線形予測子**を構築, 最尤法によってパラメーター推定する
- どういう線形予測子の構成がよいのか**モデル選択**できる; これは「あてはまりのよさ」 (deviance) と「モデルの複雑さ」 (パラメーター数) のかねあいで決まるモデル選択規準 (AIC など) の比較による

といったことを説明してきました. 今日は統計学的な検定のハナシです.

統計学における「検定」.「ふつーではない」講義を標榜してきましたが, 最終回ではこういうありがちな統計学教科書にくわしく説明されているようなハナシをすることになりました. ただしこの講義らしく, 最後まで「世の中で濫用されている『検定』とは **距離をおく**, 疑いをもちつづける, 批判的にとらえる」といったヒネくれた態度をつらぬきたい, と考えております.

「世の中で濫用されている『検定』」とはどういう言いぐさなのでしょうか? これはつまり,「考えの足りない」統計ゆーざーの態度についてであって, たとえば次のようなものです:

- 「**ゆーい差決戦主義**」: 「ゆーい差」さえあれば何を言ってもいい, 統計学的有意差は生物学的な有意差だ,「ゆーい差」がなければ「同等」だ,「ゆーい差」さえ出せるならば (あるいは消せるならば) どんな手段をもちいてもよい……などなどといった行動原理
- 「検定にかける」という表現: これも一種の「ゆーい差」万能主義みたいなものですが, 観測データについて何か discussion するときに, 何をやっているのかも理解せずに「検定にかけ」あたかもどこかのお役所で「ゆーい差」というハンコを押してもらえれば, そのデータに関してどういった議論をやってもよい, といった態度

……などなどでしょうか.

最初の回でお話ししましたように, この講義の理念としては「データをよく見てよく考えて」「統計モデルで現象を説明」というものでしたから,

理想 — この統計学授業のネライ

- 理念: スジのとおった合理的な統計解析をめざそう
- 手段: データの性質・構造によくあった手法を (データの有効利用)
- 目的: 自然現象うまく説明できるモデリングになってれば

データ解析は統計モデリング

- 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
- 基本的部品: 確率分布 (とそのパラメーター)
- データにもとづくパラメーター推定, あてはまりの良さを定量的に評価できる

「ゆーい差」なヒトたちとは対立する立場にある，といえます．では，この講義の立場からみた検定，とはどのようなものになるのでしょうか?

今日はまずモデル選択について復習し，それから「(統計学的な) 検定」についていっしょに考えていきたいと思います.

1. ポアソン回帰とモデル選択の復習

今日はモデル選択と検定について検討するために，いつもと同じような架空植物をあつかい，第 3 回の GLM (ポアソン回帰) の説明で使った種子数データ (data3a.csv ファイル) をまた再利用することにします．ただし今回は施肥処理 f_i を**完全に無視** します． [1]

1. 第 3 回の解析でみたように，施肥処理 f_i は種子数 y_i をぜんぜん説明してないし……そもそもあの架空データを作るときに**肥料は何も影響がない**と設定してポアソン乱数を生成しました.

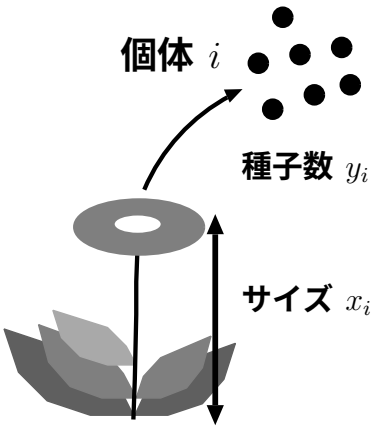


図 1: 架空植物の第 i 番目の個体 $(i = 1, 2, \cdots, 100)$

いつものごとく，CSV 形式のデータファイルを read.csv() で読みこんで， [2]

2. 今回は施肥処理をあらわす f 列は無視してください.

```
> d <- read.csv("data3a.csv")
> head(d)
   y     x f
1  6  8.31 C
2  6  9.44 C
3  6  9.50 C
4 12  9.07 C
5 10 10.16 C
6  4  8.32 C
```

このデータについては第 3 回ですでにいろいろ調べてみたので，今回はいきなりポアソン回帰，パラメーターの最尤推定値を計算してみます． [3]

3. もちろん，このあたりは第 3 回とまったく同じ結果になっています.

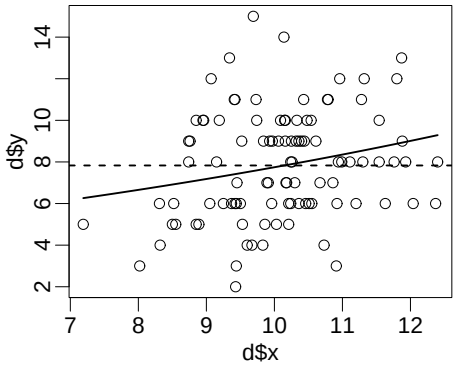
```
> (fit2 <- glm(y ~ x, data = d, family = poisson))
...(略)...
Coefficients:
(Intercept)            x
    1.29172      0.07566
...(略)...
Residual Deviance: 84.99        AIC: 474.8
```

このモデルでは個体 i の種子数平均が $\lambda_i = \exp(\beta_1 + \beta_2 x_i)$ になっていました (パラメーター数が 2 なので **モデル 2** と呼びます). 一番パラメーター数が少ない最単純モデルは $\lambda_i = \exp(\beta_1)$ と仮定するもので (パラメーターが 1 個なので **モデル 1** と呼びます),

```
> (fit1 <- glm(y ~ 1, data = d, family = poisson))
...(略)...
Coefficients:
(Intercept)
      2.058
...(略)...
Residual Deviance: 89.51        AIC: 477.3
```

となります. 両方のモデルの予測を図示してみましょう.

```
> plot(d$x, d$y) # データの表示 (施肥処理は無視)
> xx <- seq(min(d$x), max(d$x), length = 50) # 予測用の x
> abline(h = mean(d$y), lty = 2, lwd = 2) # model 1
> lines(xx, exp(1.29 + 0.0757 * xx)) # model 2
```



水平な破線がモデル 1 の予測 (サイズ x_i に依存しない), 斜めの曲線がモデル 2 の予測 (サイズ x_i に依存している) です. [4]

モデル 1 と モデル 2 の対数尤度などをまとめてみると,

4. モデル 1 の水平な線は $\exp(2.058) =$ `mean(d$y)` $= 7.83$ の位置にあります.

Model	k	$\log L^*$	**Deviance** $-2\log L^*$	**Residual deviance**	**AIC**
`Model 1`	1	-237.6	475.3	89.5	477.3
`Model 2`	2	-235.4	470.8	85.0	474.8
`FULL`	100	-192.9	385.8	0.0	585.8

となり，モデル選択規準 AIC (Akaike's information criterion)

$$\begin{aligned}\text{AIC} &= -2\times(\text{最大化対数尤度}) + 2\times(\text{モデルで使ってるパラメーター数})\\ &= (\text{Deviance}) + 2\times(\text{モデルで使ってるパラメーター数})\\ &= -2\log L^* + 2k\end{aligned}$$

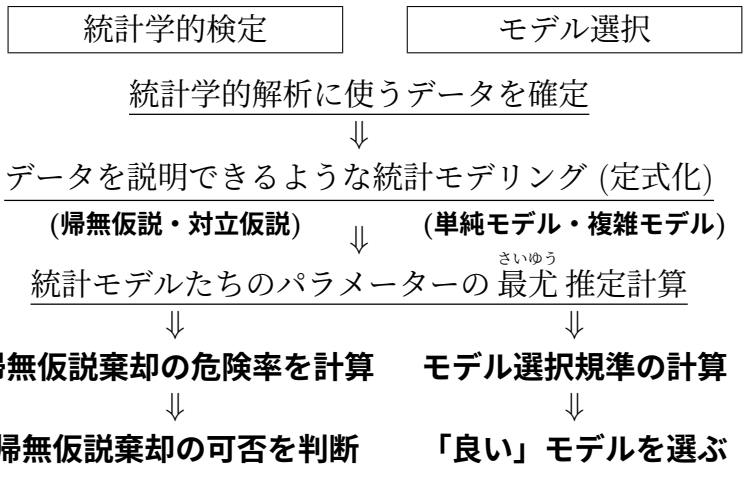
の観点からするとモデル 1 よりモデル 2 が良いモデルである，ということがわかります．[5] モデル 1 とモデル 2 の **deviance 差は 4.5 ぐらい**[6] です．

2. モデル選択と検定を比較してみよう

モデル選択はこのように「あてはまりの悪さ (deviance) の小ささ」と「モデルの複雑さ (パラメーター数)」のあいだでバランスをとろう，という方針の単純明解なものでした．

ところで生態学研究ではデータ解析というと「検定」「検定にかける」「ゆーい差」だのといったことばかりのようにあつかわれてきました．[7] ここまでこの講義であつかってきた統計モデル・尤度・最尤推定・モデル選択と「検定」はどういう関係にあるのでしょうか?

データ解析つまり統計モデリングの中でみると，この講義で示してきた統計モデリングの流れからみると，統計学的検定 [8] は下のような位置にあります．



5. なお AIC ではなく AIC_c なるものがむやみに使われている現状があるんですが……AIC_c は**データのばらつきが正規分布のときにのみ使える**モデル選択規準です (正規分布であっても使う必要ない，という議論もあります)．

6. あたりまえのことですが，これは $-2\log D^*$ で計算しても，residual deviance で計算しても同じことです．

7. さすがに近ごろは少しマシになってきましたが．

8. 統計学的な仮説検定，とでもよべばいいんでしょうかねえ…… あとで説明するように大仰な名前のわりにはたいしたコトはしていません．その点はモデル選択も同じかもしれませんが．

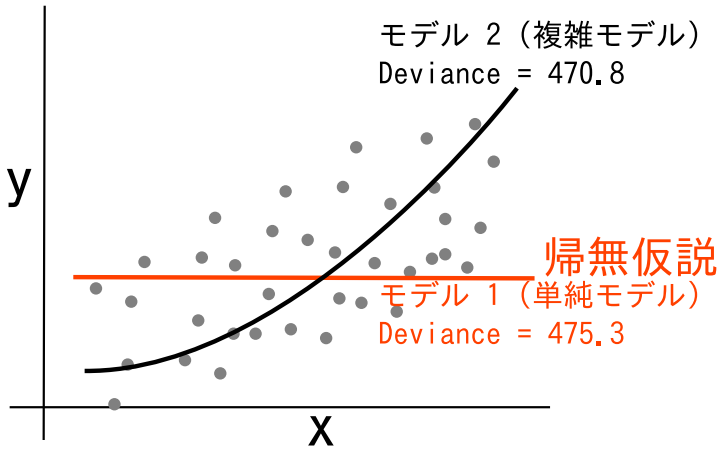
これを見ながら先ほどの架空植物種子数の例題で比較していた**モデル 1** (パラメーター 1 個だけ，サイズ x_i に依存しない) と**モデル 2** (パラメーター 2 個，サイズ x_i に依存する) の比較にあてはめて考えてみると，

- 複数のモデル (モデル 1 と 2) のパラメーターを最尤推定するところまではモデル選択も検定も同じ
- モデル選択: モデル 1 と 2 のモデル選択規準 (AIC など) を計算し，単純にそれが「良い」モデルを選ぶ (簡単!)
- 検定: …… なんだかややこしいことをやっている?

それではまず，この検定における「なんだかややこしそうなこと」について，架空植物種子数の例題のモデル 1 vs 2 比較のハナシにそって，教科書的に説明してみましょう． [9]

9. 検定のややこしさのせいでこのページは文字だらけですね.

まずそもそもモデル 1 と 2 を比較するときに，モデル 2 の「植物のサイズ x_i に依存している (個体 i の種子数平均 λ_i が x_i の関数である) といったことが言えるかどうか，がデータ解析の目的であるとおきます.



そうした場合,「サイズ x_i なんかに依存してないぢゃん」ということになってるモデル 1 (パラメーター数 1) のほうは，検定の文脈では**帰無仮説** (null hypothesis) という奇妙な名前で呼ばれることになります．モデル 2 (パラメーター数 2) は**対立仮説** (alternative hypothesis) というこれまた趣旨のわかりにくい名前になります．ともかくそういうものだということにしましょう.

この観測データのもとではモデル 1 (帰無仮説) と 2 の deviance (あてはまりの悪さ) を見たときに，パラメーター数の少ないモデル 1 が 475.3 でモ

デル 2 の 470.8 より悪くなっていました．これはどんな観測データでもいえることで，あるデータに対して確率分布 X をつかった統計モデルをあてはめるときに**パラメーター数の多いモデルのほうが必ず deviance は小さくなる**[10] ということです．

10. どんなに意味がないパラメーターであっても

モデル選択ではこのあたりに注意して，パラメーター数を増やせば deviance が小さくなるのはあたりまえなんだから，たとえば AIC は deviance だけでなく 2× パラメーター数を加えることによって，モデル複雑化の「罰則」(penalty) をあらわしていました．

3. Deviance 差を調べる尤度比検定

では，検定ではこのあたりをどう考えるのでしょうか? パラメーター数の少ないモデル 1 よりパラメーター数の多いモデル 2 のほうがあてはまりが良い (deviance が小さい) のはあたりまえだ，というところは同じです．そして deviance 差に注目します．[11] このように deviance 差に注目する検定が**尤度比検定** (likelihood ratio test) です．

11. これはモデル選択も同じなのですが．

尤度比検定とは何なのか，についてはこのあとぢりぢりと説明していきましょう．その前にちょっと「尤度比検定のごりやく」を列挙しておきましょう:

- どんな統計モデルであっても最尤推定法によってパラメーターを推定している場合に「検定」できる
  - つまりどんな確率分布を使っている場合でも使える，ということ
- モデル 1 (単純モデル) vs モデル 2 (複雑モデル) 対決のような状況[12] では最強力検定である (Neyman-Pearson の補題)
- Deviance を使うのでモデル選択との対応関係を考えやすい

12. 単純仮説の検定，ということです．

尤度比検定は尤度比

$$\frac{\text{モデル 2 の最大化尤度}}{\text{モデル 1 の最大化尤度}}$$

を比較するものですが，いつものごとく尤度ではあつかいにくいので対数尤度になおしてしかも 2 をつけた量，

$$-2\,\{\log(\text{モデル 1 の最大化尤度}) - \log(\text{モデル 2 の最大化尤度})\}$$

つまり deviance 差の大小に注目する検定です．この deviance 差のようにある検定の中で注目する量のことを**検定統計量**といいます．

4. 二種類の過誤と統計学的検定の非対称性

さて，尤度比検定の検定統計量が deviance 差である，というところまでハナシがすすみました.

そして，今回あつかっている例題では，モデル 1 (単純モデル) とモデル 2 (複雑モデル) の **deviance 差 4.5 ぐらい**となっていました.

ここで，この deviance 差について考えてみましょう．ふたつの解釈があるかと思います.

- 解釈 A: モデルを複雑化すれば (パラメーター数を増やす) とうぜんそれぐらいの差はでる，あたりまえのことだ → モデル 2 はそんなに良くない
- 解釈 B: いやいや 1 パラメーターふやしたぐらいでは，めったにこんなに増えるもんじゃない → モデル 2 はモデル 1 より良い

のどちらが正しいのか，解釈 A と B を採用したときにやってしまいそうなまちがいについて事前に考慮するのです.

↓**帰無仮説**が	モデル 1,2 の deviance 差は 「めったにない差」 (**帰無仮説**を棄却)	 「よくある差」 (棄却できない)
正しいとき	**第一種の過誤**	OK
正しくないとき	OK	**第二種の過誤**

検定ではモデル 1 よりモデル 2 のほうが良い (解釈 B)，と判断することを「帰無仮説の棄却」，その逆にモデル 1 のほうがよい (解釈 A) と判断することを「帰無仮説が棄却できなかった」とよびます. [13]

さて，ここまでのハナシとして,

- データが**モデル 1** から「ホントに」生成された場合: 「deviance 差 4.5 もあるんだからモデル 2 のほうがよい，帰無仮説は棄却だ」と考えてしまうのを**第一種の過誤** (type I error)
- データが**モデル 2** から「ホントに」生成された場合: 「deviance 差 4.5 しかないんだからモデル 2 は意味もなく複雑，モデル 1 でいいんでしょ，帰無仮説は棄却できないでしょ」と考えてしまうのを**第二種の過誤** (type II error)

13. ということで「帰無仮説の棄却」の可否がひどく重視されてるらしい，とわかります…… このあたりは「検定の非対称性」としてあとから議論します.

こんなふうに「ふたつの過誤」を同時に検討するの? とうんざりされるかもしれませんが，世の中はヒドいというかうまくできたもので，統計学的検定そのものにおいてはこのふたつのうち一方しか重視しません．[14]

統計学的検定は**第一種の過誤のみを調べる (検定の非対称性)**，つまり植物の種子数の大小に関して実際にはサイズ x_i が影響してないのに「サイズは重要だ!」などと主張してやがるかどうかの監視に専念する，ということです．

14. その理由は後述してみるつもりです…… しかしヒトコトで言えば,「両方の過誤を同時に考えるのは**めんどくさい**からイヤだ」といったところでしょうか.

5. 尤度比検定の計算の手順

「第一種の過誤だけチェック」ポリシーによってなされる統計学的検定とは具体的にはどういう手順ですすめられていくのでしょうか?

尤度比検定でモデル 1 vs 2 比較の第一種の過誤を検定してみましょう．手順はこうなります．[15]

1. まずは帰無仮説，つまりモデル 1 のような単純モデルが正しいものだと仮定する
2. つまり観測データはモデル 1 によって生成されたのだ
3. そのように生成された観測データにモデル 2 のような複雑モデルをあてはめると deviance が小さくなるのはあたりまえのことだ
4. モデル 1 と 2 の deviance 差[16] が 4.5 ぐらいというもよくあることだろう
5. ということで deviance 差が 4.5 ぐらい，あるいはそれ以上になる確率 P を計算してみよう

15. 尤度比検定ではない他の検定も同じように考えます……その場合，この説明の「deviance 差」の部分を他の検定統計量に置き換えてください.

16. deviance 差とは対数尤度差であり，すなわち尤度の比 (尤度比) を調べているということです.

このようにいわゆる P 値 (P value)[17] こと確率値が登場してきました．さてこの P 値は**第一種の過誤をおかす確率**であり，そのあつかいは,

- P 値が「大きい」: つまり deviance 差 4.5 ぐらいってのはよくあること → 帰無仮説棄却できないだろ
- P 値が「小さい」: つまり deviance 差 4.5 ぐらいってのはとても珍しいことだな → 帰無仮説を棄却しよう,「モデル 2 が正しい!」と主張してやろう

17. 「なんでも検定」を批判する論文のひとつで,「悪名たかい P value」と書いてあるのを見たような気がするなぁ……

となります．で，この「大きい」だの「小さい」だのをどうやって判断するか? すでにこれだけ検討してきて，もうあれこれと考えるのはすっかり**イヤ**になってしまったので,

- $P \geq \alpha$: 帰無仮説は棄却できない……
- $P < \alpha$: 帰無仮説は棄却，モデル 1 が無くなったので，もういっぽうの「種子数はサイズ x_i に依存」なモデル 2 で説明できる

と決めてしまうことにします．この「しきり」になる α が**有意水準**とか呼ばれてるもので，これまた「α ってどれぐらいがいいんだろう?」などとうだうだと考えるのはイヤなので 0.05 つまり 5% にしてしまう [18] ことが**なんとなく**決まっています．

まあ $\alpha = 0.05$ というのはともかく，このように特定の α を「しきい」にして帰無仮説が棄却できるかどうか，その判断を重視する統計学的検定の方式は「Neyman-Pearson のわくぐみ」とでもよぶべきもので，[19] こんにち皆さんが「検定，検定」と使いまくっている検定の基盤となるものです．

あちこちで考えるのがイヤになったおかげで，問題は整理されてきました．残された問題は「モデル 1 と 2 の deviance 差 4.5 ぐらいってのはよくあることなのか? 「モデル 1 が正しい世界」において検定統計量である deviance 差 4.5 ぐらいあるいはそれ以上になる (すなわち第一種の過誤をおかす) **確率 P ってどうやって計算**すればいいんだ?」というところまで限定されてきました．

18. つまり 20 回に一回ぐらいは第一種の過誤やってもいいぢゃん，というココロなのです……もちろんこれも根拠ふりーです．

19. Neyman-Pearson の補題はこの方式のもとでは尤度比検定が最強力検定だと示しているのですが．

6. 方法 (1) 万能なる parametric bootstrap 尤度比検定

さてさて，架空植物種子数を説明するモデル 1 とモデル 2 の deviance 差 4.5 ぐらいってのは「どれぐらいよくあることなのか」を計算する方法について説明していきましょう．今回の講義では二通りのやりかたを説明します．どちらもほぼ同じ結果になるのですが，

1. いかなるめんどうな状況でも必ず P 値が計算できる parametric bootstrap 法による尤度比検定
2. 古典的な尤度比検定 (deviance 差が χ^2 分布にしたがうと仮定)

このふたとおりを説明します．まずは parametric bootstrap 法によるものから．

まずはこの例題における deviance 差に関する補足です．架空植物の種子数の例題で，R の `glm()` による推定結果を `fit1` と `fit2` に格納しましたよね? たとえばモデル 2 の推定結果は `fit2` に入っていて，

```
> (fit2 <- glm(y ~ x, data = d, family = poisson))
...(略)...
Residual Deviance: 84.99        AIC: 474.8
```

というふうに (residual) deviance がすでに計算されていることがわかります．[20] この fit2 オブジェクトにはいろいろな情報が格納されていて，それらの情報には以下のように「ラベル」がつけられていて，

20. 同様にモデル 1 の推定結果は fit1 に入っています．

```
> sort(names(fit2))
 [1] "R"                 "aic"               "boundary"
 [4] "call"              "coefficients"      "contrasts"
 [7] "control"           "converged"         "data"
[10] "deviance"          "df.null"           "df.residual"
... (略) ...
```

たとえばこうすることで

```
> fit2$deviance
[1] 84.993
```

モデル 2 の residual deviance を取りだすことができます．これを使ってモデル 1 とモデル 2 の deviance 差をもうちょっと正確に計算しておきましょう．

```
> fit1$deviance - fit2$deviance
[1] 4.513941
```

ということで，deviance 差は 4.51 ということにしましょう．

次に，さきほど紹介した「検定のかんがえかた」のでだしを見直してみましょう．

1. まずは帰無仮説，つまりモデル 1 のような単純モデルが正しいものだと仮定する
2. つまり観測データはモデル 1 によって生成されたのだ
3. そのように生成された観測データにモデル 2 のような複雑モデルをあてはめると deviance が小さくなるのはあたりまえのことだ

ということですから，実際にポアソン乱数生成関数 rpois() を使って「モデル 1 が正しい世界」における種子数観測データを作ってみます．

```
> d$y.rnd <- rpois(100, lambda = mean(d$y))
```

ここで平均値をもともとの標本平均にしている理由は「モデル 1 が正しく，

それを使った推定値も正しい」と考えているからです．1 パラメーターつまり「全個体共通の平均だけ」モデルであるモデル 1 の推定値は標本平均が最尤推定値になっています． [21]

つぎに「モデル 1 が正しい」世界で作られた乱数データを glm() を使って 1 パラメーターモデルと 2 パラメーターモデルをあてはめてみます．

```
> fit1 <- glm(y.rnd ~ 1, data = d, family = poisson)
> fit2 <- glm(y.rnd ~ x, data = d, family = poisson)
> fit1$deviance - fit2$deviance
[1] 1.920331
```

このように,「ホントにサイズ x_i には依存してない」ポアソン乱数データに対しても

- パラメーター数 1 のモデル 1 は deviance 大 (あてはまり悪い)
- パラメーター数 2 のモデル 2 は deviance 小 (あてはまりよい)

となっていて，その deviance 差が 1.92 とわかりました．

ここまでの手つづき，すなわち，

1. 平均 mean(d$y)a のポアソン乱数を d$y.rnd に格納する
2. d$y.rnd に対するモデル 1, 2 の GLM 推定結果をそれぞれ fit1, fit2 に格納する
3. deviance 差 fit1$deviance - fit2$deviance を計算する

これによって一個の「モデル 1 が正しい世界での deviance 差」が得られます．これが parametric bootstrap 法の 1 ステップで， [22] あとはこれを 1000 回ほど繰りかえすと「検定統計量の分布」つまり「deviance 差の分布」が得られます．

このような parametric bootstrap 法を実行する関数 pb() を定義してやりましょう． [23]

21. glm() の coefficient 推定値を使って exp(2.05) としてもまったく同じですよね．

22. bootstrap 法というのはこういうふうに乱数を発生させて，それに統計モデルを適用するような統計学的な手法のことです．

23. このコードからは pb() 関数の実演中に表示されてる「いま計算中です」の点々を表示するコードは削除してます．わかりやすくするためです．ダウンロード版の pb.R にはその点々表示コードが含まれています．

```
pb <- function(d, n.bootstrap)
{
    n.sample <- nrow(d) # データ数
    y.mean <- mean(d$y) # 標本平均
    v.d.dev12 <- sapply( # PB による deviance 差の推定計算
        1:n.bootstrap,
        function(i) {
            d$y.rnd <- rpois(n.sample, lambda = y.mean)
            fit1 <- glm(y.rnd ~ 1, data = d, family = poisson)
            fit2 <- glm(y.rnd ~ x, data = d, family = poisson)
            fit1$deviance - fit2$deviance # deviance 差を返す
        }
    )
    v.d.dev12 # deviance 差 vector を返す
}
# 注: じつはこの計算のためには fit1 は不要で
#    fit2$null.deviance - fit2$deviance
# で deviance 差は計算できる (教育目的で fit1, fit2 を示している)
```

この関数定義ファイルは講義 web page から pb.R という名前でダウンロードできます.

pb.R を現在実行中の R から「見える」ディレクトリに移し,[24] source("pb.R") してやるとこのファイルを読みこみ, 関数 pb() の定義を R が記憶してくれます. あとはこれを実行すればよく, そうですね bootstrap の回数は 1000 回ぐらいとしましょう.

24. あるいは R の「ディレクトリの移動」などで R の working directory を変更.

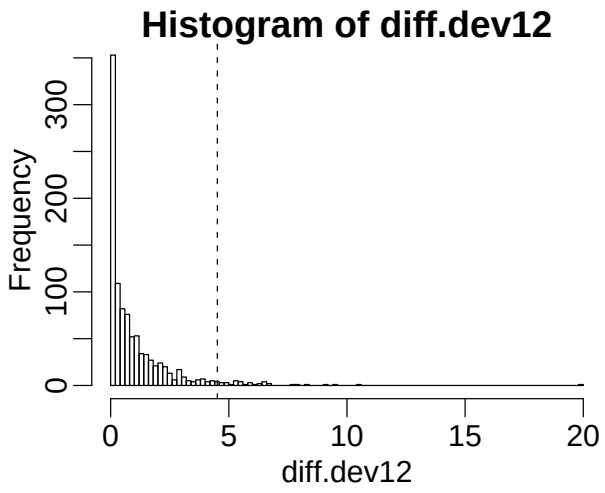
```
> source("pb.R")
> diff.dev12 <- pb(d, n.bootstrap = 1000)
# ..........................................
...(略) ...
```

乱数発生・GLM 推定計算にある程度の時間がかかって……これによって deviance 差 1000 個ぶんが diff.dev12 vector に格納されました.

```
> summary(diff.dev12)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
7.229e-08 8.879e-02 4.752e-01 1.025e+00 1.339e+00 1.987e+01
```

その値の範囲はほぼゼロから 19.9 ぐらいまで及ぶようです. ヒストグラムで「deviance 差の分布」を調べ,「deviance 差が 4.51」はどのあたりにくるのか, を調べてみましょう.

```
> hist(diff.dev12, 100)
> abline(v = 4.51, lty = 2)
```



合計 1000 個ある「deviance 差」のうちいくつぐらいが，この 4.51 より右にあるのでしょうか? 数えてみると

```
> sum(diff.dev12 >= 4.51)
[1] 38
```

ということで 1000 個中の 38 個が 4.51 より大きいことがわかりました．つまり「deviance 差が 4.51 より大きくなる確率」は 38 / 1000，すなわち $P = 0.038$ ということになります．ついでに $P = 0.05$ となる deviance 差を計算してみると，

```
> quantile(diff.dev12, 0.95)
     95%
3.953957
```

ということで 3.95 ぐらいまでは「よくある差」とみなされる，[25] ということです．検定の結論としては「deviance 差 4.51 の確率値は 0.038 だったので，[26] これは有意水準 0.05 よりも小さい」ので有意差がある (significantly different) [27] と定義され「帰無仮説 (ことパラメーター数 1 のモデル 1) は棄却され，モデル 2 が採択された」となります．

ここで紹介した **parametric bootstrap 法による尤度比検定**の特徴は次のようなものです．

- 統計モデルが明確に定義されていれば，必ず「検定統計量 [28] の分布」を乱数によって生成できて，それによって P 値が計算できる汎用性の

25. 有意水準 5% の統計学的検定のわくぐみのもとでは．

26. 尤度比検定はつねに片側検定になります…… この講義では尤度比検定しかあつかわないので，片側検定と両側検定のちがいは説明しません．

27. 「ゆーい差」とかいうとスゴいことのような印象もあるかもしれませんけど，この程度のことなのです……論文執筆では significant の使いかたに気をつけましょう．

28. ここでは deviance 差

高い方法

- すなわち，統計モデルと parametric bootstrap 法の知識さえあればどんな検定でも可能; 既存の検定が適用できない状況でも検定統計量を計算できてしまう
- ただし，ちょっとばかり計算時間を費す

つまりこれさえ知っていれば,「なんちゃら検定」のたぐいを暗記する必要がない，ということです．必要とされるのは「**いま自分はどういう統計モデルをあつかっているのか**」という認識と便利な統計ソフトウェア R だけです．

7. 方法 (2) χ^2 分布を使った古典的なる尤度比検定

前の節の parametric bootstrap 法は乱数シミュレイションによるデータ生成と最尤推定を組み合わせた統計モデルの基本に忠実な [29] 方法でした．

しかしながら，じつはこの deviance 差を比較する尤度比検定は R の中ではもっとお手軽に実施する方法があります．その手順はこうです: [30]

```
> fit1 <- glm(y ~ 1, data = d, family = poisson) # モデル 1
> fit2 <- glm(y ~ x, data = d, family = poisson) # モデル 2
> anova(fit1, fit2, test = "Chisq") # 尤度比検定
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ x
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        99     89.507
2        98     84.993  1    4.514     0.034
```

さきほどの parametric bootstrap 法ではモデル 1 と 2 のあいだの「deviance 差」が (モデル 1 が正しいという世界において) どれいぐらいになるか，を手作り関数 `pb()` を使ってランダムサンプリングして，それを `diff.dev12` vector に格納していました．

ここでは (`diff.dev12` に格納されてるような) deviance 差の確率分布にが自由度 1 [31] の χ^2 分布になるはずだ [32] という性質を利用して，その χ^2 分布 (上の例では `"Chisq"` と指定) から deviance 差が 4.51 になる P 値は 0.034 というふうに計算しています．

29. 教育的な，ともうしますか

30. `anova()` 関数の名前の由来である ANOVA とは analysis of variance つまりばらつきの解析，ここではばらつきの一種である deviance (あてはまりの悪さ) を解析しているので ANOVA の一種である analysis of deviance を実施している，というかんぢです．

31. これはモデル 1 と 2 間のパラメーター数の差です．

32. という数理統計学の定理があるのです．

これが古典的な尤度比検定のやりかたです．結論は parametric bootstrap 法の場合と同じ「deviance 差 4.51 の確率値は 0.038 だったので有意差がある，帰無仮説は棄却される」となります．

8. 検定はそんなにエラいのか?

検定についてここまで説明したことをまとめてみましょう．この例題でいうモデル 1 (単純モデル) とモデル 2 (複雑モデル) があるときに，モデル 1 のほうを帰無仮説にして「これが正しい」と仮定し，このようなわくぐみの中で，**第一種の過誤**をおかす確率である P 値を計算する，というものでした．

二つの過誤と検定の非対称性

モデル 1,2 の deviance 差は

↓帰無仮説が	「めったにない差」(帰無仮説を棄却)	「よくある差」(棄却できない)
正しいとき	第一種の過誤	OK
正しくないとき	OK	第二種の過誤

「帰無仮説が正しい」 = 「モデル 1 (単純モデル) がホントのモデル」

検定では 第一種の過誤 のみを重視する

2008–11–13 6/ 7

あらかじめ有意水準 α を 0.05 とか決めておいて，[33] $P < \alpha$ ならば

- $P < \alpha$: 帰無仮説は棄却，モデル 1 が無くなったので，もういっぽうの「種子数はサイズ x_i に依存」なモデル 2 で説明できる

と結論するものでした．では $P \geq \alpha$ だったらどうなるのでしょうか? この場合は

- $P \geq \alpha$: 帰無仮説は棄却できない……だからといって「モデル 2 よりモデル 1 がよい」とは断定できない，よくわからない [34]

と結論するのが統計学的検定の**正しい結論** ということになります.「帰無仮説が棄却できないときは帰無仮説が正しいんだ!」というハナシをときどき聞かされますが，これは統計学的な根拠が何もない**単なるまちがい**です．[35]

このように $P < \alpha$ と $P \geq \alpha$ ではずいぶんと「言えること」が違っていますね．これは検定において「第一種の過誤の危険率」 P と α の大小関係の

33. α を危険率とよぶ教科書もありますね.

34. これに対してモデル選択ではこのような非対称性はなく,「モデル 1 のほうがよい」「モデル 2 のほうがよい」と結論して問題ありません.

35. ということで「同等性の検定」とやらは完全に意味不明なんですよね. 新薬許認可のお役所てつづき用語では，統計学的同等性とは言わずに生物学的同等性とか呼んでるみたいですが.

み重視し,「第二種の過誤の危険率」 β は (たいていの「検定」ユーザーは) ぜんぜん考慮してないことが原因といえます．というか β も計算できるのですが，計算してみたところで「Neyman-Pearson の検定のわくぐみ」のもとでは何も言えません．

このような限界があるにもかかわらず,「第一種の過誤の危険率」 P 値だけを重視する検定は普及しており，何となくエラそうなかんぢです． [36] 検定は P の計算に専念することで，統計学的手法の使われかたの歴史の中で「有利さ」を発揮してきた，と私は考えています．

1. 他のことはともかく,「第一種の過誤」つまり**「いいすぎの危険性」**の確率だけは精密っぽく計算できているように見える
2. 統計的手法ユーザーの中には「とにかく帰無仮説さえ棄却できればよい」という目的のヒトもいる (むしろ多数派?) [37]
3. $P < \alpha$ かどうかの計算だけならすごくラク: 尤度比検定にしても「deviance 差 (対数尤度差) は χ^2 分布になる」という性質を利用するなら「χ^2 分布の table」みるだけで「$P = 0.05$ となる deviance 差はどれぐらいか?」がすぐにわかる

とくに 3.「計算のしやすさ」は統計学的手法がどういう順番で普及していくかを理解するカギになっていると思います．乱暴なる私見を述べれば，モデル選択をやるためにはかなり能力の高い計算機能力 [38] が必要だけれど，検定はもっと非力な計算機能力で十分，だから早くから普及し「統計といえば検定」みたいな考えがひろまったのでしょう． [39] どんなデータ解析者にとっても「検定はエラいから」検定が使われている，というわけでもなかろうということです．

36. とゆーか，実際のところは「データ解析は検定だ」とか「ゆーい差だせ，ゆーい差」とか断定したがる**すごくわかってないヒトたち**がエラそうなんでしょうなぁ……

37. これは次の節で説明してみます．

38. 1990 年代以降のパーソナルコンピューターぐらいの能力

39. そして古くから広まったのでエラそうにしている，と

9. 検定とモデル選択，それぞれの目的を考えてみる

最後に「検定とモデル選択って目的がちがうんじゃないかしらん?」ということについて考えてみましょう．

この講義は生態学など**自然科学**の研究者を対象とする統計モデリングの説明を目的としています．しかしながら，統計学的手法のユーザーは自然科学者だけではありません．むしろ応用よりな**研究・開発**なヒトたちが多いのかもしれず，たとえば，製薬会社なんかで「(製品や患者のばらつきを考慮してもなお) 旧来の薬より**新薬は性能が向上**しているか?」といったことを調べたいときには，検定はまさにうってつけのツールだと言えます．というの

も，たとえば新薬開発のハナシでいえば,「**ゆーい差さえ出ればそれでいい**, 他はどうでもいい」という立場なのでしょう: [40]

- 多くの研究時間と多額の研究費を費やした新薬は旧来薬より性能が良くて**当然**,「新薬も旧来薬も同じ」とする帰無仮説を誤棄却する確率 P がひたすら小さければそれでよい
- もし $P \geq 0.05$ となるようなら，新薬の開発は**そもそも大失敗**，検定の結論は「差があるとも何ともいえない」だろうが何だろうがどうでもいい

このような状況なら検定で第一種の過誤 P だけを重視する理由も理解できるような気がします．こういうヒトたちはじつは案外と**検定力** (または検出力; power) [41] も重視しているのかもしれません: あらかじめ「この新旧比較実験においては**これぐらいの差がでるべき**であり，こういう**実験におけるばらつき**はこのように発生するので，この差を出すために**必要な期待標本数**はこう算出でき，さらにもろもろの効果をみるために**要因をこのように統制**して……」といった難しそうな計算を事前にやっているだろう，と思います．このように事後的な検定を念頭において事前に実験の詳細を統計学的根拠にもとづいて設計するのが**実験計画法** (experimental design) です．検定の威力を十全に発揮させるためには実験計画法が必要になります．

ところで自然科学者，というか生態学 [42] みたいな野外調査において，こういう実験計画法までもちだしてデータ観測しているヒトはほとんどいません [43] ……これはまあ，そもそも実験計画法を適用しようにも「どれだけ差があればいいか」とか「対象のばらつきがこうだから差をいうために必要な標本数はこれこれ」といったことがわからない分野もある，ということなのでしょう．こういう学問は検定が使いにくい分野といえるのではないでしょうか．

また「第一種の過誤だけ重視」といった**検定の非対称性**は自然科学ではちょっと極端すぎるのかもしれません．検定の非対称性が受け入れられている [44] 理由のひとつは「『差がない』と主張する帰無仮説なんかは成立しないにきまっている，こんなダメ仮説はさっさと棄却して『差がある』と主張してやろう」なる**帰無仮説はエラくない**主義です．しかしながら，自然科学では「差がない」仮説と「差がある」仮説のエラさは同等と考えられています． [45] たとえば例題の種子数データにしても「サイズ x_i に依存していない」(つまり帰無仮説的な) モデルも「サイズ x_i に依存している」モデルも**どちらか一方がエラいといったものではない**でしょう．このようなときにはエラい・エラくない問題とは無縁なモデル選択が有効だろうと思います．

またモデル選択では「サイズという要因の影響があるかないか」というモ

40. このような「ゆーい差さえ出ればそれでいい」という極端な「ゆーい差決戦主義」を過剰におしすすめていった結果として，順位統計量にもとづく (いわゆる) ノンパラメトリック検定が誕生しました……しかしこの方式は「統計学的手法の進化系統樹」の上では先のないフクロ小路のようなもので，現代は「計算機能力向上による統計モデリング復活の時代」と私は考えています.

41. $1-\beta$ つまり第二種の過誤をやらない確率.

42. この講義はいちおう「生態学の統計モデリング」，なので……

43. ハナシによると海外にはいるそうですが……

44. とくに研究・開発なヒトたちのあいだで

45. 自然科学ではこういう「公平っぽさ」がより重視されているから，というのは理由のひとつになるかもしれません.

デル内の詳細よりも，むしろ「サイズに依存しないモデルの挙動はこうで，サイズに依存させるとこうなる」といった**統計モデルそのものの挙動**，統計モデルの予測が観測されたデータにあてはまっているのか，そのあたりを重視していると言えるのではないかと思います．[46]

46. たとえば生態学でもちいられる統計モデルは機構論的というより現象論的なものがほとんど，なわけで……現象をそれっぽく近似してるだけのモデル内である要因が作用してる・してないといったハナシはムナしくなりがちなので，そもそもそのパラメーターは統計モデル全体のふるまいをどう変えているのか，それを重視しようということです．

10. とりあえずおしまい: これからのデータ解析と統計モデリング

科学では観察・実験で得られたデータ— 構造をもった数値・記号のあつまり— をあつかいます．このとき統計学的な手法をもちいて，観察データにみられるパターンを説明できるようなうまい統計モデルを構築します．これによってデータとモデルを組みあわせて，モデルを特徴づけるパラメータなどを推定することができます．データを説明できる統計モデルを構築したい，という動機から GLM が普及してきました．これは観測データがカウントデータならポアソン分布や二項分布で現象を説明しようとするものです．

いっぽうで GLM を使っていると，何か測定できない要因が観測されたパターンに影響を与えている，ということがわかってきます．「観測されなかった/しなかった，しかし何かデータに影響をおよぼすもの」は random effects とよばれています．実際のデータ解析では，このような random effects も考慮した一般化線形混合モデル (generalized linear mixed model; GLMM) や階層ベイズモデルの応用が重要になってきます……しかし今回の統計学講義シリーズではあつかえませんでした．[47]

47. ただし「GLMM 補講」がある，というウワサも…… もちろんこれも参加する・しないは自由です．

GLMM や階層ベイズモデルの発展によって，“なかったこと” にされていた個体の差や場所の差が巧妙に統計モデル化できるようになってきました．とくに野外科学では観察できる項目が限定されるので，その観察の方針は「どういうパターンを説明したいのか? そのためには，(これまでの知見から) どの要因が重要そうであり観察すべきなのか?」をはっきりさせることがこれまで以上に重要になったと言えます．

自然科学における統計モデリングはこういった (いわゆる) random effects の影響を考慮しつつ，説明変数が興味のある現象にどう影響しているのかを明確にしていく，という方向にすすんでいくことになるでしょう．つまり，階層ベイズモデルがいよいよ広く使われるでしょう，ということです．

いろいろと説明できていない部分もありますが，今回の「生態学の統計モデリング」講義はこのあたりでいったん終了することにしましょう．皆さん，参加してくださって，どうもありがとうございます．

尤度(ゆうど) **をあつかう統計モデル**

パラメーターを確率分布として表現する Bayes 統計学

階層 Bayes モデル の MCMC 計算による推定など

最尤推定法 であつかう統計モデル

パラメーターを点推定する，random effects もあつかえる

階層ベイズモデルである一般化線形混合モデル (GLMM) 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, fixed effects のみ

最小二乗法 であつかう統計モデル

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