データ解析のための統計モデリング

全 6 回中の第 3 回 (2012–11–01 k3)

モデル選択と統計学的検定

久保拓弥 kubo@ees.hokudai.ac.jp

神戸大の集中講義 web http://goo.gl/wijx2

この講義のーとは「データ解析のための統計モデリング入門」を再編したものです

統計モデリング本 web http://goo.gl/Ufq2

もっと勉強したい人は「統計モデリング入門」を読んでね

もくじ

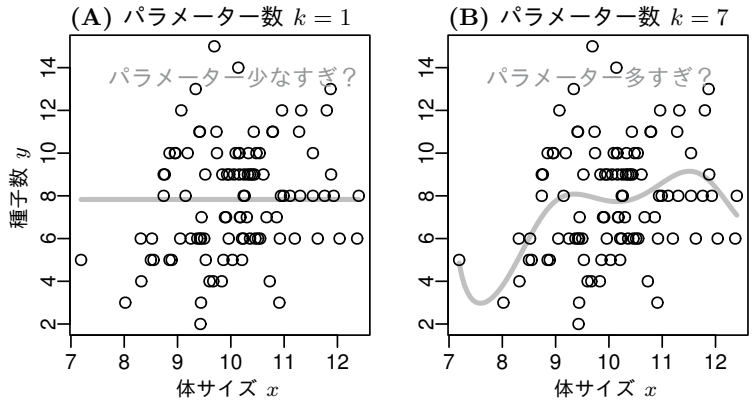

図 1　あてはまりの良さとモデルの複雑さ．前の時間 (k2) のふたつめの例題データ（**??** 節）．横軸は説明変数 x，縦軸は応答変数 y（カウントデータ）．(A) 線形予測子が切片だけの GLM ($k = 1$)．(B) 線形予測子が x の 6 次式の GLM ($k = 7$)．あてはまりを改善したいのであれば，モデルをどんどん複雑化すればよい．

この章では「良い統計モデルとは何だろう？」という疑問，あるいは「良いモデルを選びだす方法」を検討します．

複数の説明変数をいろいろと組み合わせてみて，たくさんの統計モデルが作れるようなときに，それらの中で観測データにあてはまりが良いものが,「良い」統計モデルだと考える人たちがいます[*1]．

この考えかたは正しくなさそうです．というのも，たいていの場合，複雑な統計モデルほど観測データへのあてはまりは良くなるからです．たとえば，図 1 のような例[*2] を考えてみましょう．このデータの応答変数（縦軸）はカウントデータなので，ポアソン回帰の GLM を使ってこれをうまく説明できるとします．図 1 (A) のように，線形予測子に切片だけがあるモデル（$\log\lambda = \beta_1$, パラメーター数 $k = 1$）は簡単すぎるような気もします．線形予測子を複雑化するにつれ，あてはまりは改善されていきます．しかし 図 1 (B) のように，説明変数 x の 6 次式を線形予測子とするモデル ($\log\lambda = \beta_1 + \beta_2 x + \cdots + \beta_7 x^6$, パラメーター数 $k = 7$) のようなモデルが，はたしてのぞましい統計モデルなのでしょうか？

複数の統計モデルの中から，なんらかの意味で「良い」モデルを選ぶことを**モデル選択**（model selection）といいます．モデル選択にはいろいろな方法がありますが，この章では AIC というモデル選択規準について説明します．AIC は「良い**予測**をするモデルが良いモデルである」という考えにもとづいて設計された規準です．これは「あてはまりの良さ重視」とは異なる考えかたです．

統計モデルの予測の良さとはどのようなものであり，AIC はそれをどのように評価するのでしょうか．最初に AIC の「使いかた」だけを説明するために，前の時間に登場したいろいろな GLM たちを，AIC でモデル選択してみます．モデルの複雑さとあてはまり改善の関係が，実感としてわかるのではないかと思います．

∗1　たとえば,「何でも直線回帰」な人たちが，R^2 という指標を「モデルの説明力」と信じているとか．
∗2　これは前の時間 (k2) の例題データ，架空植物 100 個体の体サイズと種子数の関係です．

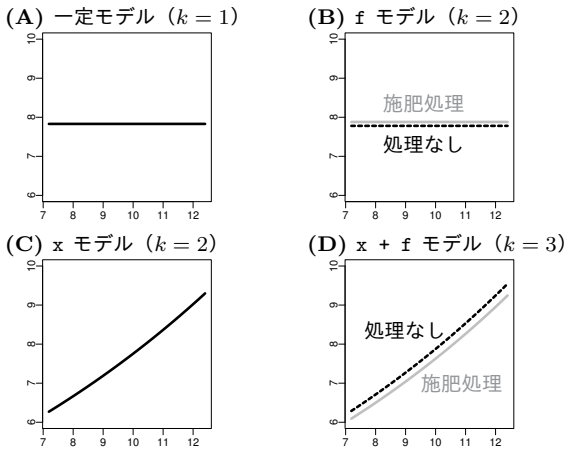

図 2　前の時間のふたつめの例題データを説明する 4 種類のポアソン回帰モデル．横軸は個体の体サイズ x，縦軸は平均種子数 λ．k は最尤推定したパラメーター数．(A) 一定モデル，説明変数不要．(B) `f` モデル: 施肥処理だけに依存．(C) `x` モデル: 体サイズだけに依存．(D) `x + f` モデル: 施肥処理と体サイズに依存．

1　データはひとつ，モデルはたくさん

前の時間のふたつめの例題では，架空植物 100 個体の種子数 y_i のデータ（**??** 節）を説明する統計モデルをいくつか作り，R の `glm()` 関数でそれぞれパラメーターを最尤推定しました．このときに，図 2 にも示しているように，同じひとつの観測データに対して，

- 体サイズ（x_i）が影響するモデル（`x` モデル; 図 2 C; **??** 項）
- 施肥効果（f_i）が影響するモデル（`f` モデル; 図 2 B; **??** 節）
- 体サイズの効果と施肥効果が影響するモデル（`x + f` モデル; 図 2 D; **??** 節）.

といった 3 種類の統計モデルをあてはめてみました．

図 2 にはもうひとつモデルが追加されていて，それは，

- 体サイズの効果も施肥効果も影響しないモデル (一定モデル; 図 2 A[*3]; 2 節)

つまり平均種子数 λ が $\exp\beta_1$ となるような「切片 β_1 だけのモデル」です．

さて，上の 4 つのモデルのうち，どれが「良い」のでしょうか？前の時間では，統計モデルのパラメーターを最尤推定しました．つまり，**対数尤度**が「いま手もとにある観測データへのあてはまりの良さ」であると考え，これを最大にするようなパラメーターの値をさがしました．

すると，あるデータを説明するいろいろな統計モデルごとに決まる，**最大対数尤度**（maximum log likelihood）つまり「あてはまりの良さ」こそがモデルの良さであると考えればよいのでしょうか？じつはそうではないだろうというのが，この章の要点です．

---

*3　図 1 (A) も前の時間のふたつめの例題データと一定モデル のくみあわせです．

2 統計モデルのあてはまりの悪さ：逸脱度

まず最初に，あてはまりの良さである最大対数尤度を変形した統計量である，**逸脱度**（deviance）[*4]について説明します（表 1）．R の `glm()` 関数を使った GLM をデータにあてはめると，推定結果には「あてはまりの悪さ」である逸脱度が出力されるので，ここで「あてはまりの良さ」との関係を整理しておきましょう．

以下では簡単のため，対数尤度 $\log L(\{\beta_j\})$ を $\log L$ と表記しましょう．この $\log L$ を最大にするパラメーターを探すのが最尤推定法です．最大対数尤度を $\log L^*$ と表記します．

逸脱度とは「あてはまりの良さ」ではなく「あてはまりの悪さ」を表現する指標で，

$$D = -2 \log L^*$$

と定義されます．これはあてはまりの良さである最大対数尤度 $\log L^*$ に -2 をかけているだけです[*5]．

平均種子数 λ_i が植物の体サイズ x_i だけに依存するモデル，$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$ を「`x` モデル」とよびます（図 2 C）．この `x` モデル の最大対数尤度 $\log L^*$ は -235.4 ぐらいでしたから逸脱度 ($D = -2 \log L^*$) は 470.8 ぐらいになります．

表 1 この節に登場するさまざまな逸脱度．$\log L^*$ は最大対数尤度．

名前	定義
逸脱度（D）	$-2 \log L^*$
最小の逸脱度	フルモデルをあてはめたときの D
残差逸脱度	D − 最小の D
最大の逸脱度	Null モデルをあてはめたときの D
Null 逸脱度	最大の D − 最小の D

しかしながら，このモデルを `glm()` であてはめると，以下のような結果が出力されます．

```
...（中略）...
Null Deviance:     89.51
Residual Deviance: 84.99        AIC: 474.8
```

この結果には，どこにも 470.8 なる数値はでてきません．そのかわり `Null Deviance`, `Residual Deviance`, あるいは `AIC` といった数量が示されています．

この結果出力に登場する逸脱度を図 3 で説明してみましょう．**残差逸脱度**（residual deviance） は，

$$D - (\text{ポアソン分布モデルで可能な最小逸脱度})$$

と定義されます．ここに登場する「ポアソン分布モデルで可能な最小逸脱度」とは何なのでしょうか？R では**フルモデル**（full model） と呼ばれているモデルの逸脱度であり，この例題ですと，**データ数が 100 個なのでパラメーター 100 個を使って**「あてはめた」モデルということです．

---

*4 deviance の訳語については前書きを参照してください．

*5 -2 をかける理由はこのあとに登場する χ^2 分布との対応関係が良くなるからです．

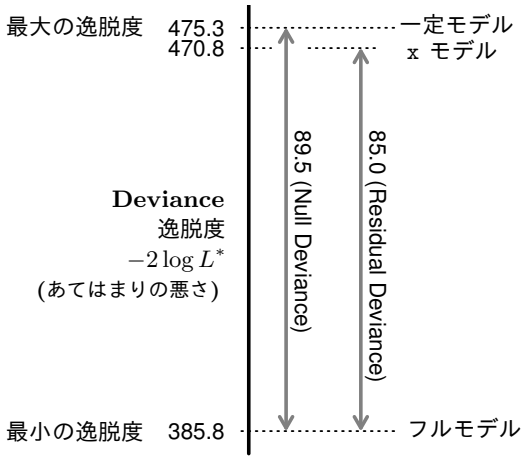

図 3 いろいろな逸脱度（deviance）．縦軸は deviance の絶対値．左右に並んでいるグレイの矢印は，それぞれ最大逸脱度 (null deviance) と残差逸脱度 (residual deviance) の値の大きさ．

最小逸脱度について説明するために，もう一度データを見なおしてみましょう．各個体の種子数 y_i は $y_i = \{6, 6, 6, 12, 10, \cdots\}$ となっていました．フルモデルとは，言ってみればこのように，100 個のデータに対して，

- $i \in \{1, 2, 3\}$ の y_i は 6 なので $\{\lambda_1, \lambda_2, \lambda_3\} = \{6, 6, 6\}$
- $i = 4$ の y_4 は 12 なので $\lambda_4 = 12$
- $i = 5$ の y_5 は 10 なので $\lambda_5 = 10$
- ...（以下略）...

100 個のパラメーターを使って「あてはめ」をする統計モデルです．

つまり，フルモデルとは全データを「読みあげている」ようなもので，統計モデルとしては価値がありません．ただし，このモデルをデータにあてはめたときに，ポアソン回帰で可能な他のどのモデルを使った場合よりも,「あてはまりの良さ」である対数尤度は大きくなり，以下のような最大対数尤度 $\log L^*$ が得られます[*6]．

```
> sum(log(dpois(d$y, lambda = d$y)))
[1] -192.8898
```

この フルモデルの逸脱度は $D = -2\log L^* = 385.8$ であり，これがこの 100 個体ぶんの観測データのもとで，ポアソン回帰で可能な最小逸脱度です．

最小逸脱度なるものが得られましたから，たとえば x モデル の残差逸脱度は，$D-$(ポアソン分布で可能な最小 D) $= 470.8 - 385.8 = 85.0$ となります．この値は，先に示した `glm()` の出力に示されていた，

```
Residual Deviance: 84.99        AIC: 474.8
```

この `Residual Deviance` と一致していますね．

このように残差逸脱度とは，このデータ解析では 385.8 を基準とする「あてはまりの悪さ」の相対値です．統

[*6] ここでは，ポアソン分布の確率を評価する R の `dpois()` 関数を使って 100 個のデータ $\{y_i\}$ に対して平均を $\{\lambda_i\} = \{y_1, y_2, y_3, \cdots\}$ とおいたときの対数尤度の和を算出しています．ポアソン分布は正規分布とは異なり，分散をゼロにできないので各 i の観測値と平均が一致していても，$\log L^* = 0$ とはなりません．

計モデルのパラメーターを多くすれば，この残差逸脱度が小さくなるらしい，ということもわかりました．

次に，図 3 を見ながら，残差逸脱度の最大値について考えてみましょう．この観測データのもとで，逸脱度が最大になるのは[*7] 「もっともあてはまりの悪いモデル」の場合です．この観測データに対するポアソン回帰の場合では，もっともパラメーター数の少ないモデル，つまり平均種子数が $\lambda_i = \exp(\beta_1)$ と指定されている，切片 β_1 だけのモデル（パラメーター数 $k = 1$）です．これは R では “null model” と呼ばれています[*8]．

このモデルでは線形予測子の構成要素が切片だけ（$\log \lambda_i = \beta_1$）なので，ここでは仮に 一定モデル （図 2 A）とします．これは R の `glm()` 関数では `glm(y ~ 1, ...)` とモデル式を指定します．

それでは，この一定モデルを使った推定計算を試みてみましょう．

```
> fit.null <- glm(formula = y ~ 1, family = poisson, data = d)
```

として推定結果を `fit.null` に格納し，その内容を表示させると，切片 β_1 の推定値は 2.06 となり，逸脱度は以下のようになりました[*9]．

```
Degrees of Freedom: 99 Total (i.e. Null);  99 Residual
Null Deviance:     89.51
Residual Deviance: 89.51        AIC: 477.3
```

このデータを使ったポアソン回帰では，残差逸脱度の最大値が 89.5 になります．これは，一定モデルの最大対数尤度が

```
> logLik(fit.null)
'log Lik.' -237.6432 (df=1)
```

となり逸脱度は 475.3 ぐらい，この逸脱度と最小 D である 385.8 の差が 89.5 ぐらいとなります．

ここまで登場したモデルについて，最尤推定したパラメーター数（k），最大対数尤度（$\log L^*$），逸脱度（$D = -2 \log L^*$），残差逸脱度（$-2 \log L^*$ − 最小 D）を表 2 にまとめてみます．パラメーター数 k さえ増やせば残差逸脱度はどんどん小さくなり[*10]，あてはまりが良くなります．

3 モデル選択規準 AIC

表 2 に示しているように，パラメーター数の多い統計モデルほど，データへのあてはまりが良くなります．しかし，それは「たまたま得られたデータへのあてはめ向上を目的とする特殊化」であり，その統計モデルの「予測の良さ」[*11]をそこなっているのかもしれません．

∗7 パラメーターを最尤推定する GLM に限定しています．

∗8 これは**帰無仮説**（null hypothesis） という別の用語を連想させますが，つながりはよくわかりません．「なんとなく帰無仮説的な，切片だけのモデル」というぐらいの意味なのでしょう．

∗9 この一定モデル の予測を図 1 (A) に示しています．

∗10 しかし `f` モデル はあてはまりの改善がものすごく小さいので，表 2 では 一定モデル とのちがいがわかりません．

∗11 **??** 節も参照してください．

表 2　種子数モデルの最大対数尤度と逸脱度．前の時間のポアソン回帰モデルの種類，最尤推定したパラメーター数 k，最大対数尤度 $\log L^*$，Deviance，Residual deviance の表．各モデルについては図 2 も参照．

モデル	k	$\log L^*$	**Deviance** $-2\log L^*$	**Residual deviance**
定数	1	-237.6	475.3	89.5
`f`	2	-237.6	475.3	89.5
`x`	2	-235.4	470.8	85.0
`x + f`	3	-235.3	470.6	84.8
フル	100	-192.9	385.8	0.0

表 3　表 2 に AIC の列を追加した．

モデル	k	$\log L^*$	**Deviance** $-2\log L^*$	**Residual deviance**	**AIC**
定数	1	-237.6	475.3	89.5	477.3
`f`	2	-237.6	475.3	89.5	479.3
`x`	2	-235.4	470.8	85.0	474.8
`x + f`	3	-235.3	470.6	84.8	476.6
フル	100	-192.9	385.8	0.0	585.8

複数の統計モデルの中から，何らかの規準で良いモデルを選択することを，**モデル選択**（model selection）と呼びます．この章では，良く使われている**モデル選択規準**（model selection criterion）のひとつ **AIC** (Akaike's information criterion) を使ったモデル選択を紹介しましょう．

AIC は統計モデルのあてはまりの良さ（goodness of fit）ではなく，**予測の良さ**（goodness of prediction）を重視するモデル選択規準です[*12]．

最尤推定したパラメーターの個数が k であるときに AIC は

$$\begin{aligned} \text{AIC} &= -2\{(\text{最大対数尤度}) - (\text{最尤推定したパラメーター数})\} \\ &= -2(\log L^* - k) \\ &= D + 2k \end{aligned}$$

と定義されます．この AIC が**一番小さいモデルが良いモデル**となります．

表 2 にさらに AIC の列を追加した表 3 を見ながら各モデルを比較すると，`x` モデル が AIC 最小の統計モデルとして選択されます[*13]．前の時間のふたつめの例題についてのモデル選択の問題は，このように解決できました．「なぜ AIC 最小のモデルが良いのか？」という疑問については「統計モデリング入門」を参照してください．

4　さてさて,「検定」のハナシですが……

データ解析において**統計学的な検定**（statistical test）はよく使われていて，統計学の教科書の中には「この場合にはこう検定して」といった解説ばかりのものもあります．「とにかく検定に帰着させればよい」と決めて

*12　他にもさまざまなモデル選択規準があります．章末に紹介している文献を参照してください．
*13　**??** 項に登場するような，ネストしている GLM のモデル選択をするときには，R の `stepAIC()` 関数を使うのが便利です．

しまえば，統計モデルといったためんどうなことを考えなくてもすむので[*14]，今後もこのような検定決戦主義は多数派でありつづけるのでしょう．

この講義は，統計モデリング試行錯誤主義とでもいうべき統計モデルによる推測・予測を重視する方向性ですから，AIC によるモデル選択と同様に，検定は推定された統計モデルを比較する方法のひとつにすぎません．この章では，どのような統計モデルでも利用可能な**尤度比検定**（likelihood ratio test） について説明します．これは前のモデル選択の章に登場した逸脱度の差に注目する考えかたです．

尤度比検定はどのような統計モデルであっても，ネストしているモデル[*15] たちを比較できます[*16][*17]．尤度比検定に限らずパラメーターを最尤推定できる統計モデルの検定を総称して，この章では，**統計モデルの検定**とよぶ場合もあります．

少しだけ用語を整理します．全パラメーターを最尤推定できる統計モデルは，**パラメトリック**（parametric）な統計モデルと総称できるかもしれません．ここでいうパラメトリックとは，比較的少数のパラメーターをもつという意味です．一部で誤用されている「正規分布を使った」という意味ではありません．また，順序統計量をつかった検定をノンパラメトリック検定[*18] とよぶ場合があります．この講義では，このような検定はあつかいません[*19]．

5　統計学的な検定のわくぐみ

最大対数尤度に注目して複数のモデルを比較するという点において，統計モデルの検定は，モデル選択と表面的には類似しているように見えます．検定とモデル選択の手順の上での共通・相違部分を図 4 にまとめてみました[*20]．

どちらの方法であっても，まず使用するデータを確定します．いったんデータを確定したら，最後までそのデータだけを使い，しかも常にすべてを使うということです[*21]．

次に目的とデータの構造に対応した適切な統計モデルを設計し，それを使ってパラメーターを最尤推定するところまでは共通です．ただしモデル選択では，パラメーターの少ないモデルと多いモデル（単純モデルと複雑モデル）とよんでいたネストしているモデルたちを，統計学的な検定ではそれぞれ**帰無仮説**（null hypothesis）・**対立仮説**（alternative hypothesis） とよびます．このあとで，検定のわくぐみの中での帰無仮説の特別あつか

∗14　与えられた観測データを好きなようにグループわけして，観測値どうしの割算値をさらに割算して何やら指標数量をあれこれこしらえて,「グループ間での指標数量の差」がゆーいになるまでこの手続きを繰り返してよいのであれば，たしかに統計モデリングなどは不要でしょうね．

∗15　**??** 節参照．

∗16　ネストしているモデルたちの比較ではない検定もあります．たとえば，パラメーターの「真の値」がわかっていて，それからずれているかどうかを調べる検定では，その「真の」モデルと推定されたモデル間の比較になります．この講義ではそのような検定はあつかいません．

∗17　さらに，この章の例題で考えているような単純な状況であれば，尤度比検定は**最強力検定**（most powerful test） です．くわしくは章末にあげている文献を参照してください．

∗18　ノンパラメトリックはこのような，順序統計量を使ったという意味だけでなく，多数のパラメーターをつかって自由自在な構造をもつ，といった意味にも使われます．

∗19　ただし一言注意すると,「正規分布じゃないから」「ノンパラは何も仮定しなくていいから」といった理由で順序統計量にもとづく検定をするのは危険です．章末の文献も参照してください．

∗20　そもそも統計モデルの検定とモデル選択が目的が異なります．それについては 10 節で述べます．

∗21　これはあたりまえのことと考えられるかもしれません．しかし学術論文の中には，個々のモデルごとに異なるデータを使って AIC を評価し，それによってモデル選択しているものもあります．検定であれモデル選択であれ，モデルごとに異なるデータを使うのは，まったくのまちがいです．

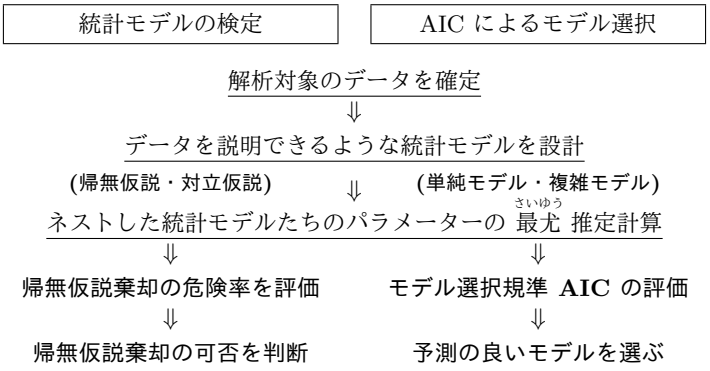

図 4　統計学的な検定とモデル選択の手順の比較.

いについて述べますが，帰無仮説とは「棄却されるための仮説」であり,「無に帰される」ときにのみ，その役割をはたす特殊な統計モデルという位置づけです.

パラメーター推定の手つづきによって推定値やあてはまりの良さが評価されたあとは，図 4 に示しているように，統計モデルの検定とモデル選択のわくぐみは異なったものになります.

モデル選択についてはすでに説明したので，ここでは統計モデルの検定の流れをおってみましょう．統計モデルの検定では,「帰無仮説は正しい」という命題が否定できるかどうか，その点だけを調べます．まず，モデルのあてはまりの良さなどを**検定統計量**（test statistic） に指定します．次に帰無仮説が「真のモデル」であると仮定して[*22]，そのときに検定統計量の理論的なばらつき（確率分布）を調べて，検定統計量の値がとりうる「ありがちな範囲」を定めます．この「ありがちな範囲」の大きさが 95% である場合は，5% の**有意水準**（significant level） を設定したといいます.

最後に対立仮説のモデルで得られた検定統計量が，この「ありがちな範囲」からはみでているかどうかを確認し，もしはみでていれば帰無仮説は棄却され，対立仮説が支持されたと結論されます[*23].

この講義ではこの検定のわくぐみを **Neyman-Pearson の検定のわくぐみ**とよぶことにします[*24]．現在よく使われている統計学的な検定の多くはこの考えかたにしたがっています.

6　尤度比検定の例題: 逸脱度の差を調べる

尤度比検定を説明するために，前の時間のポアソン回帰の例題で使った，種子数データを使います（図 5 も参照）．使用する統計モデルは，$\lambda = \exp(\beta_1 + \beta_2 x_i)$ を平均とするポアソン分布の GLM です．ネストしている一定モデルと x モデル,

- **一定モデル**: 種子数の平均 λ_i が定数であり，体サイズ x_i に依存しないモデル（傾き $\beta_2 = 0$; パラメーター数 $k = 1$）
- **x モデル**: 種子数の平均 λ_i が体サイズ x_i に依存するモデル（傾き $\beta_2 \neq 0$; パラメーター数 $k = 2$）

*22　たまたま得られた有限の観測データから推定されたモデルが「真のモデル」なんかになるのでしょうか？— このあたりも検定の作法にまつわるナゾです.
*23　ここで説明した検定のわくぐみは，より一般的な統計学的な検定においてもほぼ同じです.
*24　Neyman-Pearson ではないわくぐみの検定については，この講義ではあつかいません.

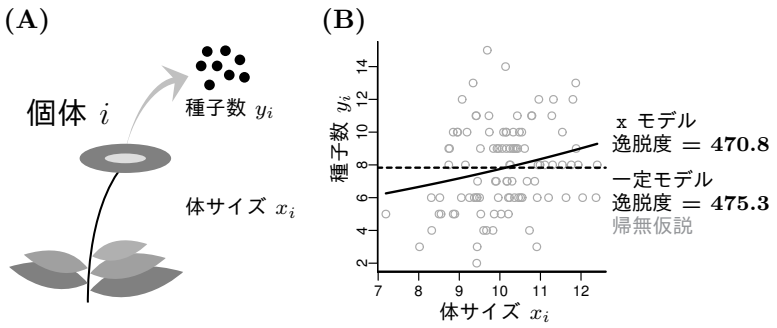

図 5　この章の例題の架空植物のデータ．(A) 前の時間のふたつめの例題と同じ．ただし施肥処理 f_i には依存しない．(B) 100 個体ぶんの観測データ（グレイの丸）と一定モデルと x モデル．水平な破線が一定モデルの予測（体サイズ x_i に依存しない単純モデル; 帰無仮説），実線の曲線が x モデルの予測（体サイズ x_i に依存している複雑モデル）．

表 4　一定モデルと x モデルの対数尤度・逸脱度・AIC．表 3 の改訂．

モデル	k	$\log L^*$	**Deviance** $-2\log L^*$	**Residual deviance**	**AIC**
定数	1	-237.6	475.3	89.5	477.3
x	2	-235.4	470.8	85.0	474.8
フル	100	-192.9	385.8	0.0	585.8

これらのうち帰無仮説となる一定モデルが棄却できるかどうかを調べます．

ポアソン回帰の結果は既出のものを編集して図 5 と表 4 に示しています[*25]．

あてはまりの悪さである逸脱度を比較すると，パラメーター数の少ない一定モデルが 475.3 で x モデル の 470.8 より悪い値になっていて，**逸脱度の差は 4.5 ぐらい**です．しかしすでに検討したように，同じデータに対してパラメーター数の多いモデルのほうが，常に逸脱度は小さくなります[*26]．

尤度比検定という名前から想像されるように，この検定では**尤度比**（likelihood ratio） というものをあつかいます．尤度比とは，たとえばこの例題の場合だと，このようになります:

$$\frac{L_1^*}{L_2^*} = \frac{\text{一定モデルの最大尤度} : \exp(-237.6)}{\text{x モデルの最大尤度} : \exp(-235.4)}$$

しかし，これをそのまま尤度比検定の検定統計量として使うわけではありません．

尤度比検定では，尤度比の対数をとり -2 をかける，つまり**逸脱度の差**

$$\Delta D_{1,2} = -2 \times (\log L_1^* - \log L_2^*)$$

に変換して[*27] 検定統計量として使います．ここで $D_1 = -2\log L_1^*$ と $D_2 = -2\log L_2^*$ とおくと，$\Delta D_{1,2} = D_1 - D_2$ ですから，$\Delta D_{1,2}$ は一定モデルと x モデルの逸脱度の差になっています．

ここでの例題データでは，一定モデルと x モデルの逸脱度の差は $\Delta D_{1,2} = 4.5$ ぐらいとなっていました．これは一定モデルに比べて x モデル ではあてはまりの悪さである逸脱度が 4.5 改善されたということです．尤度比検定では，検定統計量であるこの逸脱度の差が「4.5 ぐらいでは改善されていない」と言ってよいのかどうかを調べます．

*25　もとの図表は図 2 (A) と (C)，そして表 3.
*26　ただし,「常に」これが成立するのは，これはネストしているモデルたちを比較した場合だけです．
*27　-2 をかける理由は，標本サイズが大きい場合には，$\Delta D_{1,2}$ の分布が χ^2 分布で近似できるからです．8.2 項で説明します．

表 5　検定における二種類の過誤.

↓帰無仮説は	観察された逸脱度差 $\Delta D_{1,2}$ は「めったにない差」(帰無仮説を棄却)	観察された逸脱度差 $\Delta D_{1,2}$ は「よくある差」(棄却できない)
真のモデルである	**第一種の過誤**	(問題なし)
真のモデルではない	(問題なし)	**第二種の過誤**

7　二種類の過誤と統計学的な検定の非対称性

この章の 5 で説明したように，Neyman-Pearson の検定のわくぐみでは，比較するモデルを**帰無仮説**と**対立仮説**に分類します．この例題の場合だと，

- 帰無仮説: 一定モデル（パラメーター数 $k = 1$, $\beta_2 = 0$）
- 対立仮説: x モデル（$k = 2$, $\beta_2 \neq 0$）

と設定します[*28].

このように帰無仮説・対立仮説という概念を導入すると，一見したところで,「帰無仮説が正しくなければ対立仮説は正しい」あるいはその対偶である「対立仮説が正しくなければ帰無仮説が正しい」が成立しているかのような気がします．じつは Neyman-Pearson の検定のわくぐみでは**正しくありません**．この点については，9 節で説明するので気にしないことにして，この分類のもとで，予期される二種類の過誤を表 5 と以下に書いてみます.

- 帰無仮説が真のモデルである場合: データが**一定モデル**から生成されたのに「逸脱度の差 $\Delta D_{1,2} = 4.5$ もあるんだから x モデル（$\beta_2 \neq 0$）のほうがよい，帰無仮説は正しくない」と判断する**第一種の過誤**（type I error）
- 帰無仮説は真のモデルではない場合: データが **x モデル**から生成されたのに「$\Delta D_{1,2} = 4.5$ しかないんだから x モデルは意味もなく複雑，一定モデル（$\beta_2 = 0$）で観察されたパターンを説明できるから，帰無仮説は正しい」と判断する**第二種の過誤**（type II error）

さて，実際のところ，このような二種類の過誤をどちらも回避するのは困難です．そこで，このような二種類の過誤のうち，**第一種の過誤の検討にだけ専念**するところが，Neyman-Pearson の検定のわくぐみの要点になります.

第一種の過誤の回避に専念すればよいので，尤度比検定で必要とされる計算はずいぶんと簡単になります．この例題の場合，全体の流れは以下のようになります.

(1) まずは帰無仮説である一定モデルが正しいものだと仮定する
(2) 観測データに一定モデルをあてはめると，$\hat{\beta}_1 = 2.06$ (p.6 参照) となったので，これは真のモデルとほぼ同じと考えよう
(3) この真のモデルからデータを何度も生成し，そのたびに $\beta_2 = 0 (k = 1)$ と $\beta_2 \neq 0 (k = 2)$ のモデルをあてはめれば，たくさんの $\Delta D_{1,2}$ が得られるから，$\Delta D_{1,2}$ の分布がわかるだろう（図 6）

*28　x モデルは何が「対立」なのでしょうか？どうも対立という訳語がわかりにくいかんじで，alternative hypothesis を「代替仮説」と直訳したほうが説明しやすいように思います．なぜかというと，検定によって帰無仮説が「追放」（棄却）されたあと，現象の説明を代替するために残されたモデルであるからです.

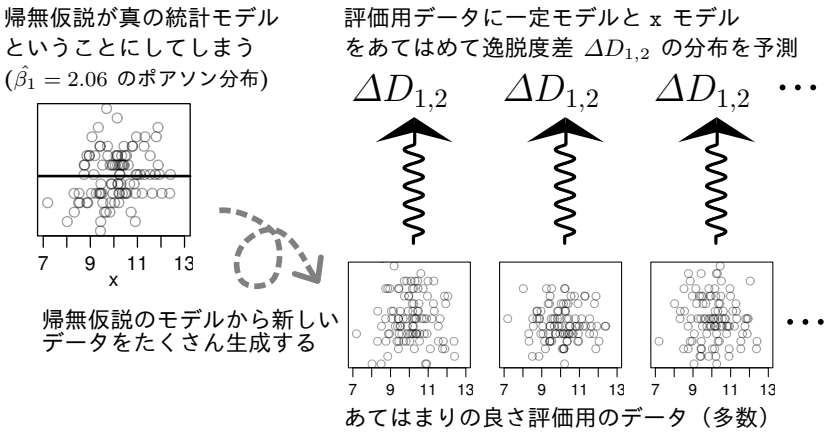

図 6　尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル （$\hat{\beta}_1 = 2.06$, p.6 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる．

(4) そうすれば，一定モデルと x モデルの逸脱度の差が $\Delta D_{1,2} \geqq 4.5$ となる確率 P が評価できるだろう

この設定のもとでの何らかの確率計算と判断によって，$\Delta D_{1,2} = 4.5$ が「ありえない」値だとみなされた場合には，帰無仮説は棄却され，残された対立仮説が自動的に採択されます．このような第一種の過誤の重視は**検定の非対称性**とよばれています[*29]．

8　帰無仮説を棄却するための有意水準

一定モデルと x モデルの逸脱度の差が $\Delta D_{1,2} \geqq 4.5$ となる確率 P は P **値**（P value）とよばれます．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」: $\Delta D_{1,2} = 4.5$ はよくあること $\rightarrow$ 帰無仮説棄却できない
- P 値が「小さい」: $\Delta D_{1,2} = 4.5$ はとても珍しいことだな $\rightarrow$ 帰無仮説を棄却しよう，残った x モデルを「正しい!」と主張してやろう

となります．

それでは，この P 値が「大きい」「小さい」はどうやって判断するのでしょうか．Neyman-Pearson の検定のわくぐみでは，**有意水準**という量 α を事前に決めておいて[*30]，以下のように判断します:

- $P \geqq \alpha$: 帰無仮説は棄却できない
- $P < \alpha$: 帰無仮説は棄却できる

ならば，この有意水準 α なる値はどのように決めればよいのでしょうか？あとは**自分で好き勝手に決める**しかありません．たとえば，$\alpha = 0.05$，つまり「めったにないこととは，20 回のうち 1 回より少ない発生件数である」といった値がよく使われています[*31]．

---

*29　第二種の過誤についての検討はどうなるのだろう，といった問題については 9 節で考えます．

*30　いかなる P 値が得られてもココロ乱されぬよう，データをとる前の段階であらかじめ有意水準 α の値を決めておくのが検定の正しいお作法とされています．

*31　なぜ 20 回に 1 回以下が「めったいにない」ことなのか，きちんとした理由は誰にも説明できません．

8.1 方法 (1) 汎用性のあるパラメトリックブートストラップ法

このあとは，P 値を評価する具体的な方法がわかれば，尤度比検定は終了します．それでは,「帰無仮説 一定モデル が真のモデルである世界」において検定統計量である $\Delta D_{1,2}$ が 4.5 より大きくなる（すなわち第一種の過誤をおかす）確率を計算する方法を考えましょう．

この章では，P 値の計算方法をふたとおり紹介します．この節では，いかなるめんどうな状況でも必ず P 値が計算できる**パラメトリックブートストラップ**（parametric bootstrap）法[*32]を説明します．次の節では，逸脱度の差が χ^2 分布にしたがうと仮定する近似的な尤度比検定を紹介します．

パラメトリックブートストラップ (PB) 法は，図 6 における「データをたくさん生成」の過程を，乱数発生のシミュレイションによって実施する方法です．以下では，R による操作を示しつつ説明してみましょう．

この章の例題データの `glm()` による推定結果は，一定モデルと `x` モデルそれぞれ `fit1` と `fit2` に格納されているとします．これら `fit1` と `fit2` オブジェクトにはいろいろな情報が格納されています．たとえば,

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

とすることで，`x` モデルの残差逸脱度を取りだすことができます．これを使って一定モデルと `x` モデルの逸脱度の差 $\Delta D_{1,2}$ を計算してみると

```
> fit1$deviance - fit2$deviance
[1] 4.5139
```

となり，やはり逸脱度の差 $\Delta D_{1,2}$ は 4.5 ということにしましょう．

統計学的な検定においては，帰無仮説が真のモデルであるとみなします．帰無仮説である一定モデルで推定された平均種子数は 7.85 個だったので[*33]，真のモデルから生成されるデータとは,「平均 7.85 の 100 個のポアソン乱数」となります．

まず，ポアソン乱数生成関数 `rpois()` を使って，真のモデルから 100 個体ぶんのデータを新しく生成してみます．

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

平均と指定している `mean(d$y)` は標本平均 7.85 です．さらに `glm()` を使って，一定モデルと `x` モデル をこの新データにあてはめてみます．

∗32 ノンパラメトリックなブートストラップ法もあります．リサンプリングや並びかえなどを使って分布を作ったり統計量を評価する方法です．

∗33 `glm()` 関数を使った推定値は，$\hat{\beta}_1 = 2.06$ ぐらいでしたから $\exp(2.06) = 7.846$ となり，これは標本平均 `mean(d$y)` とだいたい等しくなります．

```
> 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，体サイズ x_i と何の関係のない，平均値一定のポアソン乱数であるデータに対しても，逸脱度の差が 1.92 となりました.「真のモデル」である一定モデルよりも，無意味な説明変数をもつ x モデルのほうがあてはまりが良くなります.

ここまでの手順をまとめると，以下のようになります:
(1) 平均 mean(d$y) のポアソン乱数を d$y.rnd に格納する
(2) d$y.rnd に対する一定モデル,x モデルの glm() の推定結果を，それぞれ fit1, fit2 に格納する
(3) 逸脱度の差 fit1$deviance - fit2$deviance を計算する

これによって「一定モデルが真のモデルである世界」での逸脱度の差がひとつ得られます．これは PB 法の 1 ステップであり，このステップを 1000 回ほど繰りかえすと「検定統計量の分布」，この例題でいうと「逸脱度の差 $\Delta D_{1,2}$ の分布」を予測できます[*34].

この PB 法を実行するために，R の自作関数 pb() を定義してみましょう[*35].

```
get.dd <- function(d) # データの生成と逸脱度差の評価
{
  n.sample <- nrow(d) # データ数
  y.mean <- mean(d$y) # 標本平均
  d$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 # 逸脱度の差を返す
}
pb <- function(d, n.bootstrap)
{
    sapply(1:n.bootstrap, get.dd, d)
}
```

上のような関数の定義を pb.R という名前のテキストファイルに書き[*36]，R の作業ディレクトリに保存してください．R で以下のように pb.R ファイルをよみこんで，自作した pb() 関数を呼び出してみましょう.

*34 **ブートストラップ法**（bootstrap method） とは，このように乱数を使って何らかの確率分布を予測することです.
*35 じつはこの計算のためには fit1 は不要で fit2$null.deviance - fit2$deviance で逸脱度の差 $\Delta D_{1,2}$ は計算できます.
*36 自分で書かなくても **サポート web サイト** (まえがき末尾を参照) からダウンロードできます.

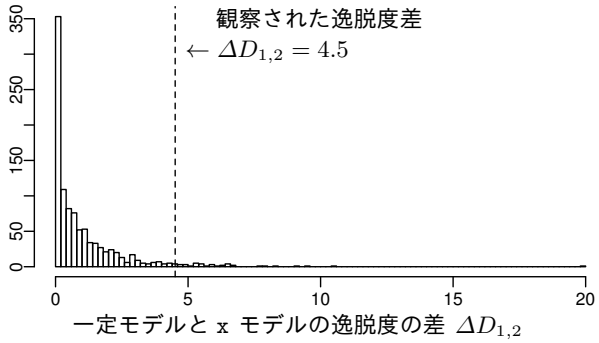

図 7 逸脱度の差 $\Delta D_{1,2}$ の確率分布．パラメトリックブートストラップ法によって生成されたヒストグラム．横軸は $\Delta D_{1,2}$．縦軸は度数（合計 1000）．縦の破線は，例題のデータに一定モデルと x モデルをあてはめて得られた $\Delta D_{1,2} = 4.5$.

```
> source("pb.R") # pb.R を読みこむ
> dd12 <- pb(d, n.bootstrap = 1000)
```

上のような R 上での操作によって，逸脱度の差 $\Delta D_{1,2}$ のサンプルが 1000 個つくられて[*37] dd12 に格納されました．その概要を summary() で調べてみましょう．

```
> summary(dd12)
     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
```

これをヒストグラムとして図示すると図 7 のようになり,「逸脱度の差 $\Delta D_{1,2}$ が 4.5」はどのあたりにくるのかもわかります.

```
> hist(dd12, 100)
> abline(v = 4.5, lty = 2)
```

合計 1000 個ある $\Delta D_{1,2}$ のうちいくつぐらいが，この 4.5 より右にあるのでしょうか？数えてみると

```
> sum(dd12 >= 4.5)
[1] 38
```

ということで 1000 個中の 38 個が 4.5 より大きいことがわかりました.「逸脱度の差が 4.5 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 $P = 0.05$ となる逸脱度の差 $\Delta D_{1,2}$[*38] を調べてみると，

*37 じつは $\Delta D_{1,2}$ のサンプル個数は 10^3 ぐらいでは十分なサイズではありません．この操作をやりなおすたびに結果がどれぐらい変わるかを調べてみてください．精度のよい結果をだすためには，n.bootstrap は 10^4 あるいはそれ以上にしたほうが良いでしょう.

*38 このような $P = \alpha$ となるような $\Delta D_{1,2}$ を棄却点 (critical point)，この値より大きい $\Delta D_{1,2}$ の領域を棄却域 (critical region または rejection region) といいます.

```
> quantile(dd12, 0.95)
     95%
3.953957
```

となり，有意水準 5% の統計学的検定のわくぐみのもとでは，$\Delta D_{1,2} \leqq 3.95$ ぐらいまでは「よくある差」とみなされます．

この尤度比検定の結論としては,「逸脱度の差 4.5 の P 値は 0.038 だったので[*39]，これは有意水準 0.05 よりも小さい」ので有意差があり（significantly different）[*40],「帰無仮説（一定モデル）は棄却され，x モデルが残るのでこれを採択」と判断します．

8.2 方法 (2) χ^2 分布を使った近似計算法

前の項で紹介した PB 法は，自分が定義した統計モデルにしたがう乱数シミュレイションによって検定統計量の分布（図 7）を生成しました．どのような統計モデルであっても，この方法を使えば近似計算なしで検定統計量の分布がわかります[*41]．

しかし，近似計算法を使うと，もっとお手軽に尤度比検定ができる場合があります．まず fit1 と fit2 に，それぞれ 一定モデル と x モデル の推定結果を格納し，

```
> fit1 <- glm(y ~ 1, data = d, family = poisson)
> fit2 <- glm(y ~ x, data = d, family = poisson)
```

以下のように anova() 関数[*42]を使います．

```
> anova(fit1, fit2, test = "Chisq")
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    4.54     0.034
```

逸脱度の差 $\Delta D_{1,2}$ の確率分布は，自由度[*43] 1 の χ^2 分布（χ^2 distribution）で近似できる場合があります．

*39　尤度比検定はつねに片側検定になります．理由はパラメーターが増えれば「観測データへのあてはまり」である最大対数尤度は必ず増大するからです（?? 節）．

*40　これは説明変数が及ぼす効果の大きさだけで決まるわけではありません．たとえばサンプルサイズが大きければ，小さな差でも統計学的には有意な差となる場合があります．また，P 値は小さければ小さいほど良いと信じている人もいますが，Neyman-Pearson の検定のわくぐみのもとでは $P < \alpha$ となっているか，なっていないかだけが問題です．

*41　より良い精度の結果を得るためには，シミュレイションのステップ数を大きくする必要があります．

*42　anova() 関数の名前の由来である ANOVA とは analysis of variance です．ただし，ここではばらつきの一種である逸脱度を調べる analysis of deviance を実施しています．

*43　これは一定モデルと x モデル間のパラメーター数の差です．

上の例では "Chisq" と指定することで，χ^2 分布近似を利用しています．結果を見ると，逸脱度の差 $\Delta D_{1,2}$ が 4.5 になる P 値は 0.034 となり，帰無仮説は棄却されます．

このようにして得られた P 値と，前の PB 法で得た $P = 0.038$ は一致していません．χ^2 分布近似は**サンプルサイズが大きいの場合に有効な近似計算**であり，この例題で調べた植物の個体数は 100 にすぎないので，このように "Chisq" 指定によって近似的に得られた P 値はあまり正確ではない可能性があります．

調査した個体数が多くない小標本のもとでは，PB 法を使って逸脱度差の分布をシミュレイションで生成するのがよいでしょう[*44]．あるいは，もしデータのばらつきがポアソン分布ではなく，**等分散正規分布**の場合には，小標本の場合の検定統計量の確率分布を利用でき，そちらのほうが χ^2 分布近似よりも正確です．たとえば，平均の差を検定統計量とする場合には t 分布，分散比を検定統計量とする場合には F 分布がよく使われています．これらの検定と尤度比検定の関係については，章末にあげている文献を参照してください．

9 「帰無仮説を棄却できない」は「差がない」ではない

この例題では，観測データを得る前にあらかじめ $\alpha = 0.05$ と定めておいて，尤度比検定の考えかたにしたがって $\Delta D_{1,2}$ の分布を予測し，その結果として $P < \alpha$ が成立して帰無仮説が棄却されたので，残された対立仮説を採択しました．

それでは，もし仮に $P \geqq \alpha$ となった場合には，どのように結論すればよいのでしょうか．その場合には，「帰無仮説は**棄却できない**（fail to reject）」と結論します．これは「帰無仮説が正しい」という意味では**ありません**．帰無仮説・対立仮説のどちらも**正しいとも正しくないともいえない**，つまり判断を保留するということです．

尤度比検定に限らず，Neyman-Pearson のわくぐみのもとでは，「帰無仮説が棄却できないときは帰無仮説が正しい」とする論法は**検定の誤用**になります．たとえば，「等分散性の検定」はよく使われていますが，これは検定の誤用です[*45]．

Neyman-Pearson のわくぐみの検定には非対称性（7 節）があるので，$P < \alpha$ となった場合と $P \geqq \alpha$ となった場合では，「結論できること」がずいぶんと異なります．

第二種の過誤の確率（表 5）を P_2 と評価することもできます[*46]．しかし，Neyman-Pearson の検定のわくぐみの中では，第一種の過誤の確率 P とは異なり，P_2 を使って何かを定量的に主張する手つづきは用意されていません．

この第二種の過誤の確率 P_2 について検討するときには，帰無仮説がまちがっていたときに棄却できる確率 $1 - P_2$ と定義される**検定力**（または**検出力**; power）がよく使われています．一般に，統計学的な検定によって帰無仮説を棄却するために実施するデータとり（実験など）では，この検定力を高めるように実験計画が定量的に設計されます[*47]．検定力を高めるためには，サンプルサイズを大きくするといった方法などがあります．

---

*44　精度を高くするために，PB 法でとりだす検定統計量の個数を十分に大きくしてください．この標本サイズは問題によるので，自分で試行錯誤して決めるしかありません．

*45　モデル選択では，「同等とするモデルが良い」「分散一定のモデルが良い」といったことが言えます．

*46　ふつうは β という記号を使いますが，この講義ではパラメーター β_1 などと混同しないように P_2 と表記します．

*47　学術誌によっては実験報告の論文に，検定力の記載を要求するところがあります．一方で，学問分野によっては，事前にも事後にも検定力などを評価せずに統計学的な検定を多用しているところもあります．

10 検定とモデル選択，そして推定された統計モデルの解釈

尤度比検定と AIC によるモデル選択は，どちらも逸脱度（あるいは最大対数尤度）という統計量に注目しています．しかし，これらふたつのモデル比較方法は，その目的とするところがまったく異なっています．モデル選択と統計学的な検定の目的のちがいに注意し，安易に混同しないように使いわける必要があります．

AIC によるモデル選択では「良い予測をするモデル」を選ぶという目的をもち,「予測の良さとは平均対数尤度」と明示したうえで，平均対数尤度を最大対数尤度とパラメーター数から推定します．

一方で，尤度比検定など Neyman-Pearson のわくぐみのもとでの統計学的な検定の目的は，帰無仮説の安全な棄却です．帰無仮説が棄却されたあとに残された対立仮説が，どのような意味で「良い」モデルなのかは明確ではありません．

自然科学の道具として使う場合に，検定にせよモデル選択にせよ,「有意でした」「このモデルの AIC が最小でした」と述べるだけで，自分の主張が正当化されるわけではありません．

統計学的な有意差とはある要因の効果の大小そのものを直接にあらわすものではないからです．なにか生物学の実験をしていて，ある実験処理の効果が「平均種子数を 1.000001 倍しか増やさない」と推定された場合であっても，サンプルサイズによっては $P < \alpha$ となったり，AIC 最小モデルとなる場合もあります．P 値は効果の大きさそのものをあらわすものではありません．

推定された統計モデルの解釈は，それぞれの研究ごとに固有なものであり，分野ごとに異なる自然現象のとらえかたに依存しているので，その文脈の中で検討すべき問題です．推定されたパラメーターはどのような値であり[*48]，標準誤差などであらわされる推定の誤差はどれほどのものなのか，それらを組みあわせたときの統計モデルの挙動はどうなると予測されるのかも示すべきでしょう．

*48 これは効果の大きさ（effect size） とよばれることもあります．