データ解析のための統計モデリング

全 6 回中の第 2 回 (2012–11–01 k2)

確率分布と一般化線形モデル (GLM)

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

「統計モデリング入門」 web http://goo.gl/Ufq2

もっと勉強したい人はこの教科書を読んでね

もくじ

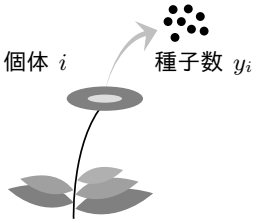

図 1　この章の例題の架空データ．架空植物の第 i 番目の個体．この植物の種子数 y_i．植物個体ごとの葉数やサイズなどについては，何のデータもない．「個体のもつ種子数をどう表現すればよいか」という単純な問題だけを検討する．

ここでは，以下のようなことを説明してみたいと思います:

- 統計学で使う**確率分布** は「正規分布」だけではない!
  - データをよくみて，それをうまく説明できるような確率分布をさがそう
- データの散布図に「セン」をひけばよし，といった考えかたはやめよう!
  - データをよくみて，よりよい「モデル」をつくろう

さてさて……いきなりハナシを始めてしまいますが……**確率分布**（probability distribution） は統計モデルの本質的な部品であり，データにみられるさまざまな「ばらつき」を表現します．この章では，このような「表現の部品としての確率分布」という考えかたを説明するために，簡単な例題データと確率分布の対応づけについて考えます．

1　ひとつめの例題: 種子数の統計モデリング

図 1 のような (まあ，何だかへんてこな) 架空の植物 50 個体からなる集団を調査していて，各個体の種子数を数えたものがデータだとします．このデータを解析しながら確率分布や統計モデルについて説明します．

この種子数データを統計モデルとして表現するために，ふさわしい確率分布は何でしょうか．まず，このデータは 0 個，1 個, 2 個，… などと数えられる**カウントデータ**（count data） です．カウントデータは非負の整数だと考えてください．このようなデータの特徴は，あとで確率分布を選ぶときの手がかりとなります．他の特徴を調べるために，このデータを統計ソフトウェア R で調べることにしましょう．

この例題データが R にすでに格納されていて[*1]，架空植物 50 個体ぶんの種子数データは `data` と名づけられているとしましょう— といった説明ではよくわからないので，実際に R を操作してみることにします．

まず，そのような状況で，R のコマンドラインインターフェイス上で `data` と入力してみましょう[*2]．

∗1　R でこの章の例題をあつかう場合には，この講義の **授業の web サイト**から `data.RData` ファイルをダウンロードしてください．R を起動し，`load(data.RData)` としてダウンロードしたデータファイルを読みこんでください．すると `data` という名前のオブジェクトをあつかえるようになります．あらかじめ，R を操作して `data.RData` ファイルが置かれているディレクトリに移動する必要があります．

∗2　これは `print(data)` という操作をしたことになります．

```
> data
 [1] 2 2 4 6 4 5 2 3 1 2 0 4 3 3 3 3 4 2 7 2 4 3 3 3 4
[26] 3 7 5 3 1 7 6 4 6 5 2 4 7 2 2 6 2 4 5 4 5 1 3 2 3
```

このように data の内容が示されます[*3]．たしかに，種子数データは 50 個の整数からなっているように見えますね．さらに， length() 関数[*4] を使えば，この data に含まれるデータ数は 50 個だと確認ができます[*5]．

```
> length(data) # data にはいくつのデータが含まれるのか？
[1] 50
```

また， summary() 関数によって，標本平均や最小値・最大値・四分位数などがわかります．

```
> summary(data) # data を要約せよ
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    2.00    3.00    3.56    4.75    7.00
```

上の summary(data) の読みかたを少し説明してみましょう．Min. と Max. はそれぞれ data 中の最小値・最大値です．また 1st Qu., Median, 3rd Qu. はそれぞれ data を小さい順にならべたときの 25%, 50%, 75% 点の値です[*6]．Median は標本中央値（中位値）と呼ばれる推定値，そして Mean は **標本平均**（sample mean）が 3.56 です[*7]．

データ解析で最も重要なのは，まず何はともあれ，そのデータをさまざまな方法で**図示してみることです**．たとえば，「種子を 5 個もつ植物は 50 個体のうち何個体だったのか？」といった度数分布を図示してみると，データ解析の役にたちそうです．

R で度数分布を得る方法はいろいろありますが，ここでは table() を使ってみましょう．

```
> table(data)
 0  1  2  3  4  5  6  7
 1  3 11 12 10  5  4  4
```

これを見ると，種子数 5 は 5 個体，種子数 6 は 4 個体といったことがわかります．これを**ヒストグラム**（histogram）[*8] として図示してみましょう[*9]．

---

*3 表示の左側の [1] だの [26] だのはそれぞれ「このすぐ右にあるデータは 1 番目のデータです」「このすぐ右にあるデータは 26 番目のデータです」を示しています．

*4 R の関数とは， length() のように名前のうしろに () がついているオブジェクトです．この () 内で**引数**（argument） をひとつまたは複数個指定します．ここでの引数は data であり， length(data) と R に指示して length() に仕事をさせることを「length() 関数を呼ぶ」と言います．

*5 R ではコメントマーク # から行末までは読みとばされます．つまり，その部分に注意や説明をメモできます．

*6 これらの値が四分位数です．R が出力するこれらの値は観測値そのものではなく，たいていの場合，補間予測された値です．

*7 このように標本を加減乗除して得られる統計量の名前にはいちいち「標本 ～」と明記すべきなのですが，この講義ではそのような厳密さは守られていません．

*8 度数分布図と呼ばれることもあります．

*9 seq(-0.5, 9.5, 1) は $\{-0.5, 0.5, 1.5, 2.5, \cdots, 8.5, 9.5\}$ という数列を生成します．関数 hist() の breaks 引数にこのような列を与えると，「-0.5 より大 0.5 以下」「0.5 より大 1.5 以下」— といった区間ごとのヒストグラムを作図します．

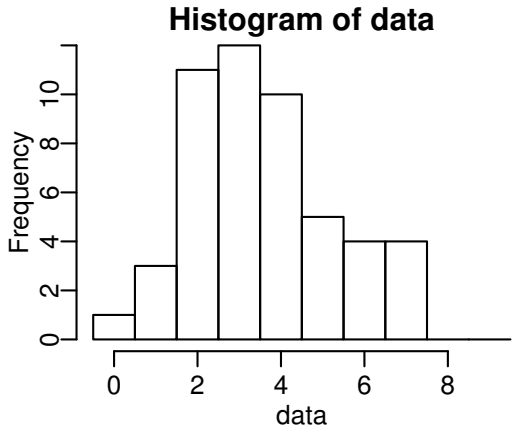

図 2　例題の種子数データのヒストグラム（度数分布図）．横軸は種子数，縦軸は架空植物の個体数．全個体数は 50．R の `hist()` 関数による図示．

```
> hist(data, breaks = seq(-0.5, 9.5, 1))
```

R では上のように指示すれば，図 2 に示しているようなヒストグラムがえられます．

あるデータの**ばらつき**（variability） をあらわす標本統計量の例として，**標本分散**（sample variance） があげられます．

```
> var(data)
[1] 2.9861
```

また，**標本標準偏差**（sample standard deviation） とは標本分散の平方根です．R ではこんなふうに計算できます．

```
> sd(data)
[1] 1.7280
> sqrt(var(data))
[1] 1.7280
```

2　データと確率分布の対応関係をながめる

ここまで植物個体ごとの種子数データを調べてみて，以下のような特徴があることがわかりました．

- 1 個， 2 個， … と数えられるカウントデータ
- 1 個体の種子数の標本平均は 3.56 個
- 個体ごとの種子数にばらつきがあり，ヒストグラムを描くとひと山の分布になる

さて，データに見られるこのようなばらつきをあらわすためには，**確率分布**（probability distribution） という考えかたを使います．この講義では確率分布についての抽象的な検討から始めるのではなく，この例題デー

タを解析するために必要な確率分布をまず導入し，その性質を説明していきます．

この章の例題である種子数データを統計モデルとして表現するためには，とりあえず**ポアソン分布**（Poisson distribution） とよばれる確率分布が便利である— ということにしておきましょう[*10]．

確率分布とは**確率変数**（random variable） の値とそれが出現する確率を対応させたものです．この例題にあてはめて言うと，ある植物個体 i の種子数 y_i のように，個体ごとにばらつく変数が確率変数です．そして，この確率変数 y_i はたとえば $y_i = 2$ といった値をとるのだとすると，そのように $y_i = 2$ となる確率はどれぐらいなのか，というところに興味があります．

確率変数の値とそのとりうる確率の対応づけ，たとえばこの例題でいうと「個体 1 の種子数 $y_1 = 2$ となる確率はどれぐらいか？」を表現する確率分布は（後述するように）比較的簡単な数式で定義され，**パラメーター**（parameter） の値に依存して「分布のカタチ」が変わります．

例題データにそって，さらに具体的に説明してみましょう．ポアソン分布で指定できるパラメーターはひとつだけであり，それは分布の平均です．得られたデータとポアソン分布の対応関係とは何なのか，いまだによくわかりませんが— とりあえず，この例題のデータの標本平均は 3.56 でしたから，ここからしばらくは，平均 3.56 のポアソン分布とはどのようなものなのかを調べてみましょう．

数式を使ったポアソン分布の定義などは次の節以降であつかうことにして，ここでは「平均 3.56 のポアソン分布」なるものを，R を使ってグラフとして図示しましょう．平均 3.56 のポアソン分布にしたがって「種子数が y であると観察される確率」を生成させるには，たとえば以下のように R に指示します．

```
> y <- 0:9
> prob <- dpois(y, lambda = 3.56)
```

R では `dpois(y, lambda = 3.56)` と関数を呼びだすと，`prob` オブジェクトに「ある個体の種子数が y 個である確率」が格納されます．それを表形式で出力すると図 3 のようになります．

これではわかりにくいので図示してみると，

```
> plot(y, prob, type = "b", lty = 2)
```

このように `plot()` 関数を呼ぶと，図 4 のように種子数 `y` とその確率 `prob` の関係が示されます．

これらの図表に示している平均 3.56 のポアソン分布とは何なのでしょうか？ここでは，種子数の個体間のばらつきをあらわす近似的な表現手法として，このポアソン分布を導入しています．図 3 や図 4 に示されているように，ある植物個体の種子数がゼロである確率は 0.03 ぐらい，一番確率が高くなるのは一個体に 3 個の種子をもつ場合で，その確率は 0.21 ぐらい，といったことを表現したければ，平均 3.56 のポアソン分布を持ちだせば良いのではないか，というアイデアです．

統計モデリングにおいてはこのように確率分布を使えば，ばらつきのある事象・現象を記述できるとみなします．つまり，観測データのヒストグラム（図 2）に平均 3.56 のポアソン分布（図 4）を重ねたときに，図 5

*10　ここでは，とりあえずカウントデータを近似的に表現するために，便利でお手軽な確率分布としてポアソン分布を使っていると考えてください．この場合は，とくに何か生物現象のメカニズムからポアソン分布が導出されたわけではありません．実際のデータではポアソン分布だけを使った統計モデルではうまく説明できない場合がほとんどです．この問題は第 2 日目であつかいます．

```
   y       prob
1  0 0.02843882
2  1 0.10124222
3  2 0.18021114
4  3 0.21385056
5  4 0.19032700
6  5 0.13551282
7  6 0.08040427
8  7 0.04089132
9  8 0.01819664
10 9 0.00719778
```

図 3　平均 3.56 のポアソン分布の確率分布，$y \in \{0,\ 1,\ 2,\ \cdots,\ 9\}$．本文のように R 内で種子数 y と各種子数が観察される確率 $p(y \mid \lambda)$ を計算して prob に格納しておき，さらにコマンドプロンプトで cbind(y, prob) と指示すると得られる出力．

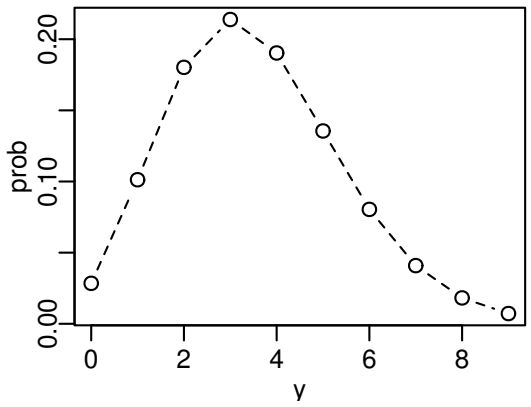

図 4　平均 $\lambda = 3.56$ のポアソン分布．種子数 y とその確率 prob の関係が示されている．図 3 の表を図にしたもの．R の plot() 関数の引数，type = "b" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

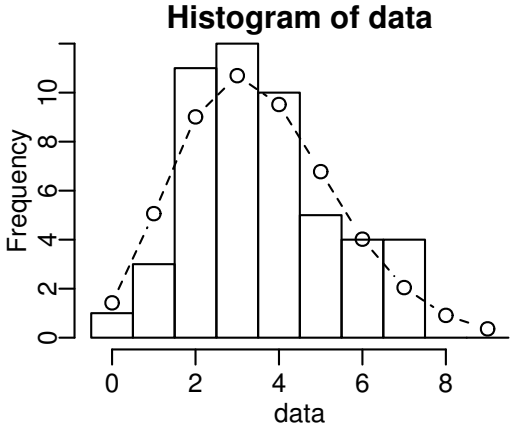

図 5　観測データと確率分布の対応をながめる．ヒストグラムは図 2 と同じ．それに重ねられている丸と破線は y 個の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

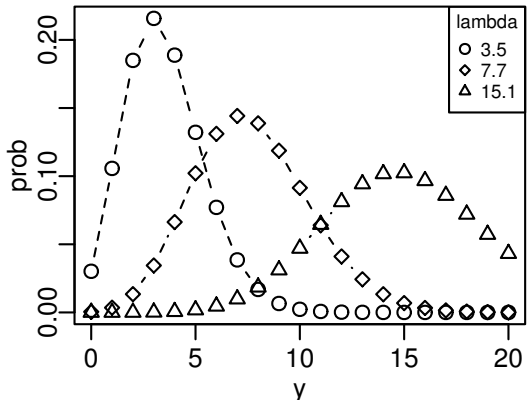

図 6　さまざまな平均（λ）のポアソン分布．$\lambda \in \{3.5,\ 7.7,\ 15.1\}$.

のような結果[*11] が得られたのであれば，観察されたばらつきがポアソン分布で表現できているみたいだなぁと考えます．

それでは次に，この「図の見た目」による納得気分をもう少し定量的に示す方法を検討します．とくに興味があるのは，どのような確率分布，あるいは統計モデルを使って与えられた観測データを説明できるのか，確率分布のパラメーターはどう決めればよいのか，そして観測データを説明できる良い統計モデルとは何か，といった問題です．

3　ポアソン分布とは何か？

この章の例題の観測データと確率分布の対応関係を検討するために，前の節で何の説明もなく登場したポアソン分布について，もう少しくわしく説明します．

ポアソン分布の確率分布 は以下のように定義されます．

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

この $p(y \mid \lambda)$ は平均が λ（ラムダ）であるときに，ポアソン分布にしたがう確率変数が y という値になる確率です[*12]．$y!$ は y の階乗で，たとえば 4! は $1 \times 2 \times 3 \times 4$ をあらわしています．

平均 λ はポアソン分布の唯一のパラメーターです．上の定義を使って図 3 や図 4 の確率が評価されています．この他のポアソン分布の性質をあげます．

- $y \in \{0,\ 1,\ 2,\ \cdots,\ \infty\}$ の値をとり，すべての y について和をとると 1 になる

$$\sum_{y=0}^{\infty} p(y \mid \lambda) = 1$$

- 確率分布の平均は λ である（$\lambda \geqq 0$）

*11　図 5 のような図を作るためには，まず `hist()` 関数でヒストグラムを作図し，次に予測を描画するために `lines(y, 50 * prob)` と指示する必要があります．

*12　教科書によっては，この関数 $p(y \mid \lambda)$ を確率関数あるいは確率質量関数とよんでいます．

• 分散と平均は等しい: $\lambda =$ 平均 $=$ 分散

ポアソン分布のパラメーター λ を変化させると，確率分布は図 6 に示しているように変化します．

この講義では，この例題の種子数データのようなタイプのデータは，ポアソン分布あるいはポアソン分布と他の分布を混ぜた確率分布で統計モデル化します．なぜ，この種子数のばらついている様子を記述する統計モデルの部品として，このポアソン分布が選ばれたのでしょうか．そのように尋ねられたら，以下のような理由を挙げてみればよいでしょう．

(1) データに含まれている値 y_i が $\{0,\ 1,\ 2,\ \cdots\}$ といった非負の整数である（カウントデータである）
(2) y_i に下限（ゼロ）はあるみたいだけど上限はよくわからない
(3) この観測データでは平均と分散がだいたい等しい[*13]

この講義でいうばらつきとは，統計学用語でいうところの**誤差**（error） のことです．誤差とは**測定誤差**（measurement error） のことだと考えられがちなのですが，統計学用語でいう誤差とはもっと広い意味をもつものです．

たとえば，この例題のように全 50 個体の架空植物の種子数がひとつの（つまり全個体で共通する一個の平均 λ をもつ）ポアソン分布にしたがうとしましょう．このときに，図 2 などに見られるようなばらつきは，調査している人間が数えまちがえたために，個体によって種子数が異なるのだ— とは**考えていません**[*14]．そうではなく，個体ごとの種子数が同一の確率分布にしたがっている場合であっても，なんらかの理由[*15] で個体ごとに異なる種子数になっていると考え，その不確定性も「誤差」とよばれます．この講義ではまぎらわしいので，可能なかぎり誤差という語は使わず，ばらつきとよびます．

この例題のように，λ という全個体共通である平均のポアソン分布から，ばらつきのある種子データが発生したという統計モデルを作るためには，その他にもいくつかの前提が必要です．たとえば，植物のサイズなどいろいろな条件をそろえている状況では，どの個体でも種子数の平均が同じであり，個体ごとにちがいはないと考えています．

このように説明しても理解できない人もいるので，もう少しだけしつこく記述します．上でのべていることは，

• 個体 1 の種子数は平均 λ のポアソン分布にしたがうと仮定する $\rightarrow$ 観測された種子数は 2 だった
• 個体 2 の種子数は平均 λ のポアソン分布にしたがうと仮定する $\rightarrow$ 観測された種子数は 2 だった
• 個体 3 の種子数は平均 λ のポアソン分布にしたがうと仮定する $\rightarrow$ 観測された種子数は 4 だった
• —（以下，同様）—

という意味で，このように仮定すると全 50 個体のデータから全個体に共通する λ は 3.56 ぐらいではないかなぁといった憶測が可能になる — ということです．

このほかにも，植物個体どうしは独立であり，個体間の相関や相互作用はないといった前提も必要です[*16]．

---

*13　分散は 2.99 ぐらいと評価されています（ 節）．このような条件が満たされない場合の統計モデリングについては，明日の講義で登場します．
*14　もちろん人間の数えまちがいも含めた統計モデルも構築できますが，この講義ではあつかいません．
*15　この「なんらかの理由」をきちん特定した上で，それに対応するような確率分布を選べるのか，と問いつめられると— 必ずしも可能でないと答えざるをえません．このあとの 6 節では，データの外形的な特徴だけをみて確率分布を選び，そこからの逸脱があれば複数の確率分布をまぜて統計モデリングする，といった話をしています．
*16　現実のデータではこのような条件の成立は疑わしいのですが，統計モデルの説明を簡単にするために，今日の説明ではそういうことにしておきます．

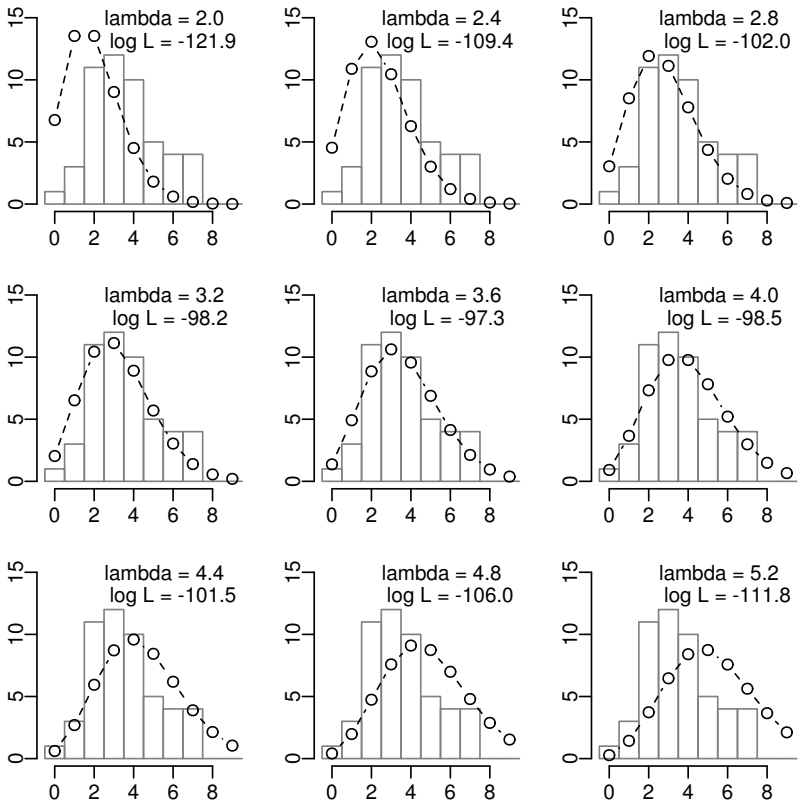

図 7　平均 λ（`lambda`）を変化させていったポアソン分布と，観測データへのあてはまりの良さ（対数尤度 `logL`）．すべてのヒストグラムは図 2 と同じ．

4　ポアソン分布のパラメーターの最尤推定

それでは，今度は確率分布のパラメーターを，観測データにもとづいて推定する方法について考えてみましょう．ここでは **最尤**（さいゆう） **推定**（maximum likelihood estimation）あるいは最尤推定法というパラメーター推定の方法を紹介します[*17]．これはポアソン分布だろうが正規分布だろうが，どのような確率分布を使った統計モデルにも適用できます[*18]．

最尤推定について説明する前に，表記方法について少し整理してみましょう．この章では，第 i 個体の種子数の観測値は y_i とします．つまり $\{y_1,\ y_2,\ y_3,\ \cdots,\ y_{49},\ y_{50}\} = \{2,\ 2,\ 4, \cdots,\ 2,\ 3\}$ ということです．あるいはまた 50 個まとめて一文字で表記するときには $\{y_i\}$ あるいは $\mathbf{Y} = \{y_i\}$ と書きます．また，すでにポアソン分布の確率分布の定義にもあらわれましたが，$p(y_i \,|\, \lambda)$ は平均 λ が決まっているポアソン分布において，y_i という値が発生する確率です．たとえば，図 3 に記されているように，$\lambda = 3.56$ のときは $p(y_1 = 2 \,|\, \lambda = 3.56)$ は 0.180 となります．

さてさて，最尤推定法は **尤度**（ゆうど） なる「あてはまりの良さ」をあらわす統計量を最大にするようなパラメーター（この例題では λ）の値を探そうとするパラメーター推定方法です．尤度の実態は，ある λ の値を決めたときに，

---

*17　「尤も」の訓読みは「もっとも」であり，もっともらしいのもっともです.「最尤」はもっともももっともらしい.

*18　最小二乗法は，データのばらつきが正規分布であるとした統計モデルの最尤推定法と同等です.

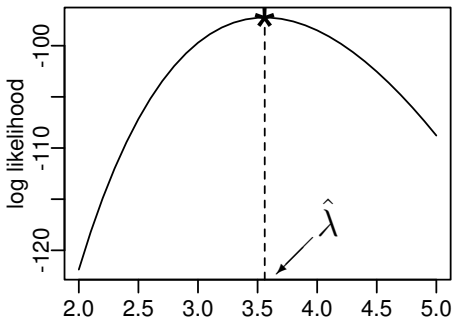

図 8　この章の例題の観測データ（植物 50 個体の種子数）のもとでの λ（横軸）と対数尤度の関係．$\lambda = 3.56$ で対数尤度が最大になるので，これが最尤推定値 $\hat{\lambda}$ となる．図 7 の λ を連続的に変化させた場合に対応している．

すべての個体 i についての $p(y_i \mid \lambda)$ の積です．たとえば，いまデータが 3 個体ぶん，たとえば，$\{y_1,\ y_2,\ y_3\} = \{2,\ 2,\ 4\}$，これだけだった場合に，図 3 をにらみながら計算してみると，尤度はだいたい $0.180 \times 0.180 \times 0.19 = 0.006156$ といった値になります．

尤度はパラメーターの関数なので $L(\lambda)$ と書きます．この例題の場合には，このように定義されます

$$\begin{aligned} L(\lambda) &= (y_1\text{が } 2 \text{ である確率}) \times (y_2\text{が } 2 \text{ である確率}) \\ &\quad \times \cdots \times (y_{50}\text{が } 3 \text{ である確率}) \\ &= p(y_1 \mid \lambda) \times p(y_2 \mid \lambda) \times p(y_3 \mid \lambda) \times \cdots \times p(y_{50} \mid \lambda) \\ &= \prod_i p(y_i \mid \lambda) = \prod_i \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!}, \end{aligned}$$

となります[*19]．なぜ積になるのかといえば，「y_1が 2 である」かつ「y_2が 2 である」かつ— と 50 個の事象が同時に真である確率を計算したいからです[*20]．

この尤度関数 $L(\lambda)$ はそのままではあつかいにくいので，対数変換した**対数尤度関数**（log likelihood function）を使ってパラメーターを最尤推定します．

$$\log L(\lambda) = \sum_i \left(y_i \log \lambda - \lambda - \sum_k^{y_i} \log k \right)$$

まずは，平均をあらわすパラメーター λ を変化させていったときに，ポアソン分布のカタチと対数尤度がどのように変化するのかを調べてみましょう（図 7）．

図 7 を見ると，対数尤度が大きい（ゼロに近い）ほど観測データとポアソン分布が「似ている」ように見えます．

さらに，対数尤度 $\log L(\lambda)$ と λ の関係を調べるために，図 8 のような図を作ってみましょう．この作図のための R コードは以下のようになります．

---

*19　$\prod_i$ はこの場合，$i \in \{1,\ 2,\ \cdots,\ 50\}$ についての積です．**??** 節も参照．

*20　たとえば，おもて・うらが等確率で出現するコイン 50 枚を投げたときに，「全部おもて」になる確率は 0.5^{50} ですよね．

```
> logL <- function(m) sum(dpois(data, m, log = TRUE))
> lambda <- seq(2, 5, 0.1)
> plot(lambda, sapply(lambda, logL), type = "l")
```

この例題の場合，3.5 から 3.6 あたりで対数尤度が最大になる λ が見つかりそうです．なお，対数尤度 $\log L$ は尤度 L の単調増加関数なので，対数尤度が最大になる λ において尤度も最大になります．

対数尤度が最大になる λ を $\hat{\lambda}$ としましょう[*21]．図 8 に示されているように，対数尤度関数は最大値で対数尤度関数の傾きがゼロとなる λ を探しだせばよいので，さきほど登場した対数尤度関数 $\log L(\lambda)$ をパラメーター λ で偏微分して，

$$\frac{\partial \log L(\lambda)}{\partial \lambda} = \sum_i \left\{ \frac{y_i}{\lambda} - 1 \right\} = \frac{1}{\lambda} \sum_i y_i - 50$$

これがゼロである場合，

$$\hat{\lambda} = \frac{1}{50} \sum_i y_i = \frac{\text{全部の } y_i \text{の和}}{\text{データ数}} = \text{データの標本平均} = 3.56$$

となります．最尤推定値 $\hat{\lambda}$ は 3.56 であり，この単純な例題の場合，最尤推定値は標本平均に等しくなります．

このように対数尤度あるいは尤度が最大になる $\hat{\lambda}$ を**最尤推定量**（maximum likelihood estimator），さらに $\{y_1,\ y_2,\ y_3,\ \cdots,\ y_{49},\ y_{50}\} = \{2,\ 2,\ 4, \cdots,\ 2,\ 3\}$ というふうに具体的な y_i の値を使って評価された $\hat{\lambda} = 3.56$ のことを**最尤推定値**（maximum likelihood estimate） とよびます．

尤度と最尤推定について少し一般化してみましょう．たとえば，θ をパラメーターとする確率分布から観測データ y_i が発生した場合，その確率を $p(y|\theta)$ とすると，尤度は

$$L(\theta|\mathbf{Y}) = \prod_i p(y_i|\theta),$$

で対数尤度は

$$\log L(\theta|\mathbf{Y}) = \sum_i \log p(y_i|\theta),$$

となります．最尤推定とは，この対数尤度を最大にするような $\hat{\theta}$ を探しだすことです．この考えかたは確率分布 $p(y|\theta)$ がポアソン分布でない場合であっても，同じように適用できます[*22]．

実際のデータ解析で使う統計モデルはもっと複雑なものになるので，こんなに簡単に最尤推定量は導出できません．そこで計算機を使って最尤推定値に近い値を探しだします．たとえば，次の章以降で紹介する一般化線形モデルの推定関数などでは，数値的な試行錯誤によって対数尤度ができるだけ大きくなるようなパラメーターを探しだします．

5 統計モデルの要点: 乱数発生・推定・予測

ここで確率分布が統計モデリング・データ解析の中で果たす役割について整理してみましょう．私たちが何か観測データ，たとえばこの章の例題の種子数のデータである，

*21 $\hat{\lambda}$ はラムダハットと読みます．この講義ではハットは推定値（ただしこのあとしばらくは推定量）をあらわします．
*22 正規分布のような連続値の確率密度関数の尤度方程式の例は，あとでちょっと紹介するかもしれません．

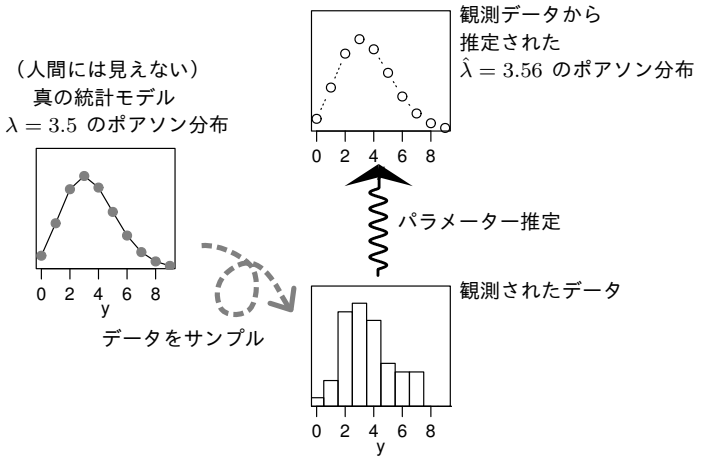

図 9　統計学における推定．自然（データの発生もと）を確率分布であると見たてて，得られたデータをそこから発生した乱数のセットだと考える．限定された個数のデータから，もとの確率分布に近い確率分布を探すのが「推定」.

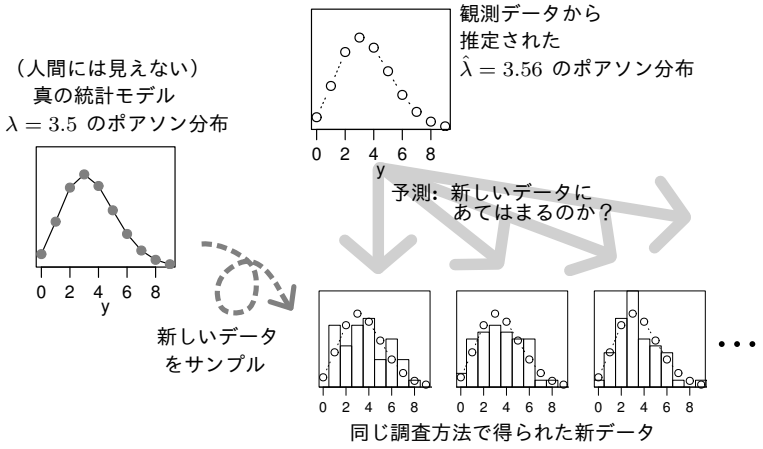

図 10　統計学における予測．図 9 で推定されたモデルが，次に発生するデータの分布を予測する．予測の良さは新データへのあてはまりで**検証**（validation） できる.

```
> data
 [1] 2 2 4 6 4 5 2 3 1 2 0 4 3 3 3 3 4 2 7 2 4 3 3 3 4
[26] 3 7 5 3 1 7 6 4 6 5 2 4 7 2 2 6 2 4 5 4 5 1 3 2 3
```

このような数字の羅列を見たときに,「こういうばらつきのあるデータは，何か確率分布から発生したと考えればあつかいやすそうだなぁ」と考えることが統計モデリングの第一歩となります.

このときにデータ解析者のアタマの中では，図 9 のように考えています．まずデータを発生させた統計モデルが「真の統計モデル」（真のモデル）であり，これが平均 3.5 のポアソン分布であるとしています[*23]．統計モデルの中の確率分布を使って**乱数**（random number） を発生させることができます．これを**サンプリング**

*23　この事例では，観測者には見えない真のモデルがたまたまポアソン分布であり，観測者は「データがカウントデータだから」という理由でポアソン分布を部品とする統計モデルを使っています.

（sampling） とよぶこともあります．この講義では，このようにサンプルされた乱数のあつまり（標本）が，観測データであると考えています．

この観測データを見たときに，データのばらつきは「まあポアソン分布で説明できるだろう」と仮定したとしましょう．このときに「パラメーター λ はどんな値？」という問いに答えるのが**推定**（estimation），あるいは（モデルのデータへの）**あてはめ**（fitting） といいます．図 9 では，最尤推定によって $\hat{\lambda} = 3.56$ 得られました．つまり，平均 3.56 のポアソン分布はデータに最もあてはまりが良いと言えます．

観測データと統計モデルの対応づけでもうひとつ重要なのは**予測**（prediction） です．推定で得られた統計モデルを使って，同じ調査方法で得られる**次のデータ**の分布などをみつもることが予測です．

統計モデルを使った予測にはいろいろなものがあります．

- 次に得られる説明変数の平均だけを示す[*24]
- 平均だけでなく，次に得られるデータはこの範囲にちらばるはずという**予測区間**（prediction interval） も示す（例: 図 5）

たとえば時系列構造をあつかっている統計モデルの予測であれば，いわゆる将来予測となりますし，空間構造のあるデータなどで**欠測データ**（missing data） をうめるのも予測の一種です[*25]．

統計モデルの良さを評価するときに，**予測の良さ**（goodness of prediction） という考えかたが重要です[*26]．これは推定されたモデルが新しく得られたデータにどれぐらい良くあてはまるかをあらわします．たとえば，図 10 では，追加の観測によって真のモデルから新しく得られたデータたちに，図 9 で推定された平均 3.56 のポアソン分布をあてはめてみて，あてはまりの良さ（対数尤度）を評価しようとしています[*27]．

5.1 データ解析における推定・予測の役割

推定と予測を比較してみましょう．統計モデリングの初心者からすると推定に比べて，予測は難しいと感じるかもしれません．たとえばこの講義で紹介しているような例題ならば，R の推定関数の使いかたさえ知っていれば推定は簡単にできます．

一方で，推定されたモデルを使って予測をするためには，自分が使っている統計モデルをよく理解していなければなりません[*28]．統計モデルを推定したら，そこで解析を終了するのではなく，その**推定結果をうまく図示**することが重要であり，このような図示はいろいろな場合において，推定されたモデルを使った予測になっています[*29]．

科学で使われるモデルの良さとは，そのモデルの予測の良さによって決まります．また，推定されたモデルによる予測を試みることで，自分が使っている近似的・現象論的な統計モデルの理解が深まり，またその不備が判明することもあります．このような点もふくめて，推定したモデルによる予測は重要です．推定しただけ，あるいは「検定」しただけでは十分とは言えません．

---

*24 いわゆる「回帰の線をひく」なども含まれます．
*25 このような欠測データについては授業の最後のほうで紹介します．
*26 次の時間では，予測の良さにもとづくモデル選択として検討します．
*27 あてはまりの良さの結果は図・本文どちらにも示していません．
*28 ただし次の時間に紹介する予定なのですが，R では予測のための,「あまり考えなくても使える」関数がいくつか準備されています．
*29 次の章以降の例題でも推定結果を図示していますので，それがどのような意味で予測になっているか考えてみてください．

6 確率分布の選びかた

ある観測データを解析する統計モデリングにおいてまず考えるべき点は,「この現象がどのような確率分布で説明されそうか」ということです．この章の例題からはいったん離れて，多種多様な確率分布がある中で，自分のデータの統計モデルに使えそうな分布の選びかたを考えてみましょう．とりあえずデータをみたら次の点に注意してみてください．

- 説明したい量は**離散**か**連続**か？
- 説明したい量の**範囲**は？
- 説明したい量の標本分散と標本平均の関係は？

この講義では，カウントデータの統計モデルで使う確率分布として，以下のふたつを使います:

- **ポアソン分布**（Poisson distribution）: データが離散値，ゼロ以上の範囲，上限とくになし，平均 ≈ 分散
- **二項分布**（binomial distribution）: データが離散値，ゼロ以上で有限の範囲（$\{0,\ 1,\ 2,\ \cdots,\ N\}$），分散は平均の関数

これらはそれぞれ，この講義で紹介する統計モデルの中の重要な役割をはたします．また，連続確率分布である

- **正規分布**（normal distribution）: データが連続値，範囲が $[-\infty, +\infty]$，分散は平均とは無関係に決まる
- **ガンマ分布**（gamma distribution）: データが連続値，範囲が $[0, +\infty]$，分散は平均の関数

これらについては，今回の授業ではあまりくわしく説明しません．

6.1 もっと複雑な確率分布が必要か？

現実のデータ解析では,「このデータのヒストグラムは複雑に見えるので（あるいはカウントデータなのにポアソン分布のように見えないので），ポアソン分布とか簡単な確率分布では表現できないかもしれない」と考えたくなる状況に頻繁に遭遇します．世の中には多種多様な確率分布があるので，その中から複雑なものを選んで利用すればよいのでしょうか？この講義では，そんなに多種類の確率分布を使いこなさなくても，現実に見られる多彩なばらつきを，統計モデリングの工夫次第であつかっています．

たとえば，いろいろと条件をそろえて観測した種子数のカウントデータであるのに，どうもポアソン分布には見えないんだ，となる場合もあるでしょう．このような問題は，個体たちは均質ではない（たとえば，遺伝的に均質でない，生育環境にわずかな差があった）にもかかわらず，観測者がそのデータをもっていない場合に発生します．このようなときには,「データ化されていない個体差・見なかった個体差」をくみこんだ統計モデリングが必要となります．これについては，明日の授業で紹介します．

7 ふたつめの例題: 個体ごとに平均種子数が異なる場合

この章の例題は，ひとつめの例題と良く似ていますが，個体の大きさがさまざまであったり，あるいは植物たちをふたつのグループにわけて，それぞれで異なる実験処理をほどこしていたりする点が異なっています．

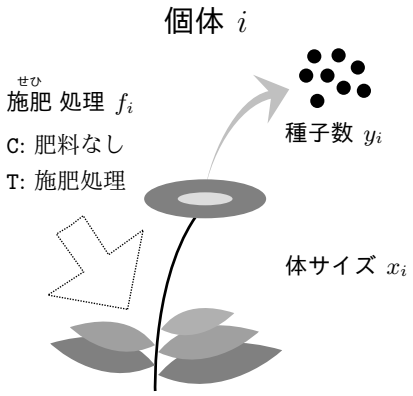

図 11　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

図 11 に示しているような架空植物 100 個体を調査して得られた，個体ごとの種子数データがあるとしましょう．これを統計モデルを使ってどのように表現すればよいか，というのがこの章でとりくむ問題です．植物個体 i の種子数は y_i 個であり，また個体の属性のひとつである**体サイズ**（body size） x_i が観測されています[*30]．この体サイズは植物の大きさの大小をあらわす正の実数です．体サイズの大きさが種子数に影響しているかもしれないので，考えられているので，その効果を調べるために x_i も測定されたのだとしましょう．

さらに，全個体のうち 50 個体（$i \in \{1, 2, \cdots, 50\}$）は何も処理をしていない（処理 C，いわゆるコントロール）けれど，残り 50 個体（$i \in \{51, 52, \cdots, 100\}$）には肥料を加える処理（施肥 処理, 処理 T）をほどこします．体サイズ x_i とは無関係に施肥処理がなされたとします．

個体ごとに異なる属性は，x_i や f_i といった観測データとして与えられています．これらは観測・設定された「個体ごとのちがい」であることに注意してください．

8　観測されたデータの概要を調べる

まずは，この架空植物の種子数の観測データを R で取りあつかう方法を簡単に紹介してみましょう．この方法は，他の章の例題のデータに対しても，同じように適用できます．

この章のデータは CSV ファイル[*31] という形式で保存されていて， **授業の web サイト** からダウンロードできます．データのファイル名は `data3a.csv` であるとしましょう．R では

```
> d <- read.csv("data3a.csv")
```

と命じるだけでファイルを読みこんで，その内容を格納したデータフレイムに `d` という名前が付けられます．このデータフレイムというデータ構造は，とりあえず「表 (table) のようにあつかえるデータ構造」であると

*30　ここでいう植物の体サイズは何でもいいのですが，たとえば，植物個体の高さだと考えてもらってかまいません．実際の植物のデータ解析ではもっと別の量で体サイズを表現することが多いのですが，この講義ではそのあたりはいいかげんにしておきます．

*31　CSV とは comma-separated value のことで，表の各要素がコンマで区切られたフォーマットのことです．これは汎用性のあるデータフォーマットで，たいていのスプレッドシートソフトウェアなどで CSV 形式を指定してファイルにデータを保存できます．

考えてください．R のコマンドプロンプトで d あるいは print(d) とすると，全 100 個体ぶんのデータがディスプレイ上に表示されます[*32]．

```
> d
    y     x  f
1   6  8.31  C
2   6  9.44  C
3   6  9.50  C
...（中略）...
99  7 10.86  T
100 9  9.97  T
```

このように d というオブジェクトには，全 100 個体ぶんのデータがあたかも 100 行 3 列の行列のような形式で格納されているように見えます．このデータでは，最初の列 y には種子数，x には個体の体サイズ，f には施肥処理の値が入っています．

この d の列ごとにデータを表示させてみましょう．x と y 列は以下のように表示されます．

```
> d$x
  [1]  8.31  9.44  9.50  9.07 10.16  8.32 10.61 10.06
  [9]  9.93 10.43 10.36 10.15 10.92  8.85  9.42 11.11
...（中略）...
 [97]  8.52 10.24 10.86  9.97

> d$y
  [1]  6  6  6 12 10  4  9  9  9 11  6 10  6 10 11  8
 [17]  3  8  5  5  4 11  5 10  6  6  7  9  3 10  2  9
...（中略）...
 [97]  6  8  7  9
```

また施肥処理の有無をあらわす f 列はちょっと様子がちがっています．

```
> d$f
  [1] C C C C C C C C C C C C C C C C C C C C C C C C C
 [26] C C C C C C C C C C C C C C C C C C C C C C C C C
 [51] T T T T T T T T T T T T T T T T T T T T T T T T T
 [76] T T T T T T T T T T T T T T T T T T T T T T T T T
Levels: C T
```

このように表示される理由は，f の列には **因子**（factor） というクラスのデータが格納されているためです．この講義では仮に因子型とよびます．R の read.csv() 関数は，CSV 形式のデータファイル内に C だの T だのといった文字を含む列を見つけたときには，これらを factor に変換します．このように変換された f 列は C と T の 2 **水準**（level） からなる値で構成されていて，上の R 出力では Levels の行で f 列内の水準を示しています．因子型の水準の値には順番があり，この場合は C が 1 番目で T が 2 番目となっています．これは read.csv() 関数が「水準の順番はアルファベット順」というルールで変換したためです[*33]．

R の class() 関数を使うと，あるデータオブジェクトがどういう型（正確にはクラス）に属しているかを

*32 データフレイムを表示させる関数は print() だけではありません．たとえば， head(d) とすると最初の 6 行だけが表示，head(d, 10) とすると最初の 10 行だけが表示されます．

*33 もちろんユーザーが指示してこの水準の順番を変更できます．

調べられます.

```
> class(d) # d は data.frame クラス
[1] "data.frame"
> class(d$y) # y 列は整数だけの integer クラス
[1] "integer"
> class(d$x) # x 列は実数も含むので numeric クラス
[1] "numeric"
> class(d$f) # そして f 列は factor クラス
[1] "factor"
```

さて，R の summary() 関数を使って，この d と名づけられたデータフレイムの概要を調べてみましょう.

```
> summary(d)
       y               x           f
 Min.   : 2.00   Min.   : 7.190   C:50
 1st Qu.: 6.00   1st Qu.: 9.428   T:50
 Median : 8.00   Median :10.155
 Mean   : 7.83   Mean   :10.089
 3rd Qu.:10.00   3rd Qu.:10.685
 Max.   :15.00   Max.   :12.400
```

データフレイムの summary() はこのように列ごとの要約が表示されます．どのように “summary” されるのかは，列の型に依存しています．数値型 (numeric) である y と x 列の要約については，ひとつめの例題で説明しました．因子型である f 列の表示は，C が 50 個，T が 50 個ありますよ，と要約されています.

9 統計モデリングの前にデータを図示する

統計モデリングにとりくむときに，そのデータを「いろいろな図にしてよく見る」点は何度でも強調しておきたいところです．前の節で紹介した summary() による統計量表示だけでなく，この節で紹介するような作図関数を使ってデータのばらつきかたを視覚的に把握するようにしましょう[*34].

観測データに余計な手を加えないで，データ全体をよく見るには plot() 関数などを使うと便利でしょう.

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```

これによって描かれた図 12 は，横軸に x 列， 縦軸に y 列をとった**散布図**（scatter plot） です.

横軸が因子型であっても，上と同じように plot(df, dy) と指定すれば，自動的に図 13 のような**箱ひげ図**（box-whisker plot） が生成されます[*35].

これらの図を見て，なんとなくわかることは,

*34 作図するときには，できるだけデータそのままを使いましょう．データ列どうしの割り算値などをでっちあげて作図するのは，しばしばまちがいのモトになります.

*35 R が図 13 のような箱ひげ表示をデフォルトにしている理由は「標本分布をよくみろ」といったことを勧めているためなのかもしれません．箱ひげ図は分布のゆがみなども図示されるので，よく見かける「平均 ± 標準偏差」だけの図示よりすぐれています.

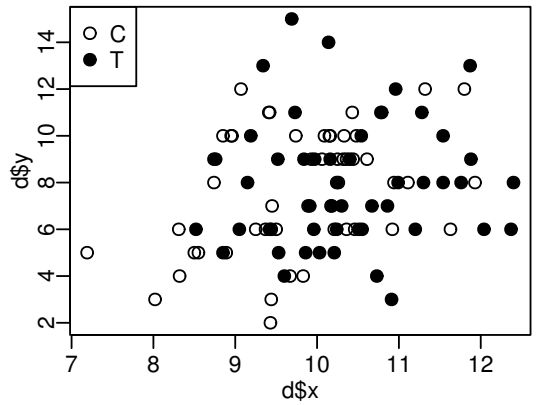

図 12　例題の架空データの図示．植物の種子数 y_i と，体サイズ x_i や施肥処理 f_i の関係を示している．白丸は施肥処理なし（処理 `C`），黒丸は施肥処理あり（処理 `T`）．

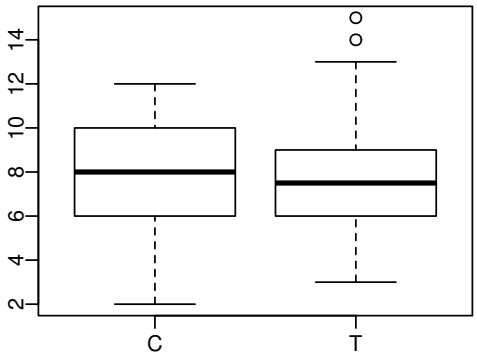

図 13　植物の種子数の分布を，施肥処理 f_i でグループわけした箱ひげ図（ `plot(d$f, d$y)` の出力）．ハコの上中下の水平線はそれぞれ 75%, 50%, 25% 点，上下のヒゲの範囲が近似的な 95% 区間，マルはその近似的 95% 区間からはみだしたデータ点をあらわしている．理由は不明だが，このように作図すると軸ラベルはつかない．ユーザーが指定して軸ラベルをつけることは可能．

- 図 12 を見ると，体サイズ x が増加するにつれ種子数 y が増えているように見えるけれど，あまりはっきりしない
- 肥料の効果 f はぜんぜん無いように見える

といったところでしょうか.

10 ポアソン回帰の統計モデル

それでは，このようなカウントデータである種子数データをうまく表現できそうな統計モデルを作ってみましょう．この例題についても，ポアソン分布を使った統計モデルでデータのばらつきを表現できそうです．前の章では，平均種子数 λ が全個体で共通の値であると仮定しました．しかし，この章の例題では，個体ごとの平均種子数 λ_i が体サイズ x や施肥処理 f に影響されるようなモデルを設計します.

まず最初に個体 i の体サイズ x_i だけに依存する統計モデルについて考えてみます[*36]．説明変数は x_i であり，応答変数は種子数 y_i です．施肥効果 f_i はあまり種子数に影響がなさそうなので，ここではひとまず無視します.

ある個体 i において種子数が y_i である確率 $p(y_i \mid \lambda_i)$ はポアソン分布にしたがっていて，

$$p(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

と仮定します．ここまではひとつめの例題のモデルと同じです.

10.1 線形予測子と対数リンク関数

この個体ごとに異なる平均 λ_i を説明変数 x_i の関数として定義しなければなりません．ここでは，ある個体 i の平均種子数 λ_i が，

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

であるとしてみましょう．この講義では，β_1 や β_2 を**パラメーター**（parameter）と呼び，β_1 を**切片**（intercept），β_2 を**傾き**（slope） とよぶことにします[*37]．数式ではよくわからないので，平均種子数 $\lambda_i = \exp(\beta_1 + \beta_2 x_i)$ の関係を図示してみると，図 14 のようになります[*38].

このような定式化にどのような意味があるのかを検討する前に，ここで**線形予測子**（linear predictor） と**リンク関数**（link function） という，GLM を特徴づけるふたつの概念を紹介しておきましょう．このモデルの平均種子数 λ_i の式は，

$$\log \lambda_i = \beta_1 + \beta_2 x_i$$

*36 この統計モデルでは説明変数である体サイズ x_i という観測値には測定時の誤差がまったくない，と仮定しています．じつは，このような「説明変数の誤差」を無視すると推定結果に偏りが生じる場合があります．このような問題をあつかいたい場合は，明日説明するベイズモデルの技法を使って，応答変数 y_i だけでなく説明変数 x_i に関する統計モデル化が必要になります．しかしながら，この講義ではこの方法については説明していません.

*37 これらのパラメーターを**係数**（coefficient），説明変数 x_i を**共変量**（covariate） とよぶ場合もあります.

*38 応答変数 y_i の種類によっては，サイズ x_i をそのまま線形予測子に加えるのではなく，**サイズの対数** $\log x_i$ を加えて平均が $\exp(\beta_1 + \beta_2 \log x_i)$ とするほうが適切なモデルとなる場合もあります．これは，サイズと平均のあいだにアロメトリックな関係 $\log(\text{平均}) = \beta_1 + \beta_2 \log x_i$ を仮定していることになり，$x_i \to 0$ のときに平均もゼロに近づきます.

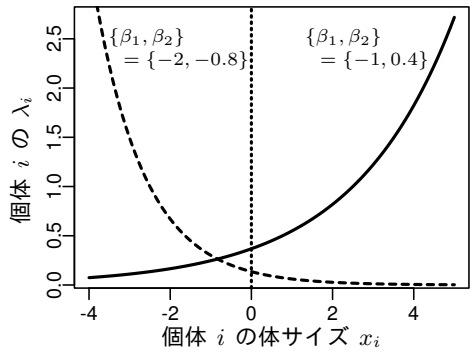

図 14　個体 i の平均種子数 λ_i と体サイズ x_i の関係．$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$ と設定している．体サイズ x_i が負の値になるのは奇妙ではあるけれど，このモデルは x_i が 7 から 13 ぐらいの範囲だけで適用できる近似だと考えることにしよう．

と変形できます．このときに右辺 $\beta_1 + \beta_2 x_i$ は線形予測子とよばれます．これがもし $\beta_1 + \beta_2 x_i + \beta_3 x_i^2$ であったとしても線形予測子とよばれます．その理由はこの式が $\{\beta_1, \beta_2, \beta_3\}$ の線形結合になっているからです．また，上の式は $\log \lambda_i =$（線形予測子）となっていますが，このように（λ_iの関数）=（線形予測子）となっている場合，左辺の「関数」は**リンク関数**とよばれます．この場合は対数関数が指定されていますから，リンク関数は **対数リンク関数**（log link function）とよばれます．ポアソン回帰をする場合，たいていはこの対数リンク関数を使用します．

この講義では，ポアソン回帰の GLM には対数リンク関数，明日説明するロジスティック回帰の GLM にはロジットリンク関数を使用しています．これらは数学的に都合のよい性質があるので，ポアソン分布・二項分布の **正準リンク関数**（canonical link function）と呼ばれています．R の `glm()` では，特に指定しなければ 各 `family`（ばらつきの確率分布）ごとに異なる正準リンク関数が使用されます．

ポアソン回帰の GLM で対数リンク関数を使う理由は，これが「推定計算に都合よく」かつ「わかりやすい」からです．推定計算に都合がよいのは，$\lambda_i = \exp($線形予測子$) \geqq 0$ となっているところです（図 14）．ポアソン分布の平均は非負でなければなりません．このように対数リンク関数を使うと，説明変数やパラメーターがどのような値になってもこの条件が守られるので，R に最尤推定値を探索させるときに便利です．

また，対数リンク関数が「わかりやすい」と述べた理由は，要因の効果が積であらわされるからです．この点については，12 節で検討します．

10.2　あてはめとあてはまりの良さ

ポアソン回帰とは，観測データに対するポアソン分布を使った統計モデルの**あてはめ**（fitting） であり，この統計モデルの対数尤度 $\log L$ が最大になるパラメーター $\hat{\beta}_1$ と $\hat{\beta}_2$ の推定値を決めることです．データ $\mathbf{Y}$ のもとでの，このモデルの対数尤度は，

$$\log L(\beta_1, \beta_2) = \sum_i \log \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

となります．線形予測子は $\log \lambda_i = \beta_1 + \beta_2 x_i$ となっているので，λ_i が β_1 と β_2 の関数であることに注意してください．

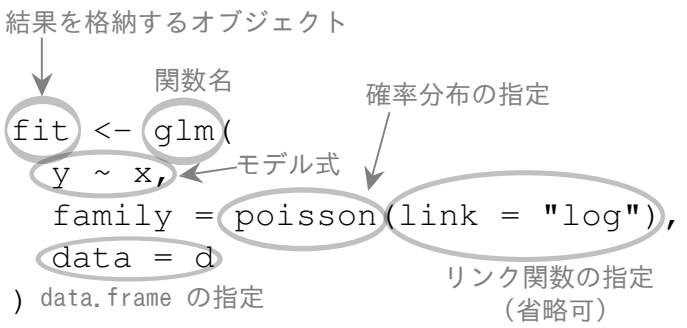

図 15 glm() 関数の引数の指定方法.

今回の統計モデルでは，複数のパラメーター $\{\beta_1, \beta_2\}$ を同時にあつかうので，最尤推定量の導出は簡単ではありません．しかしながら，実際のポアソン回帰では，たいていの場合，数値的な試行錯誤によって最尤推定値を探しだすので，推定量が解析的に導出できなくても問題ありません．

R ではたいへんお手軽に GLM のあてはめができるようになっていて，

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

と指定すれば[*39]，切片 β_1 と傾き β_2 の最尤推定値が得られます． glm() 関数で指定している内容は図 15 のようになります．family = poisson は「分布はポアソン分布を使ってね」という指示です[*40]．ここでは，得られた推定結果などを fit と名づけられたオブジェクトに格納するように命令しています[*41]．

それでは，この例題の推定結果を格納している fit を調べてみましょう．まず概要を表示してみます[*42]．

```
> fit # あるいは print(fit) としてもよい
Call:  glm(formula = y ~ x, family = poisson, data = d)

Coefficients:
(Intercept)            x
    1.29172      0.07566
...（以下略）...
```

summary(fit) 関数を使うと，さらに詳細な結果が表示されます．ここでは，パラメーターの推定値だけを抜粋します．

---

*39 関数内での値の指定を**引数**（argument） とよび，たとえば data = d という指定では，d が引数であり data は仮引数とよばれます．最初の y ~ x という指定では仮引数を省略しています．省略せずに書くと formula = y ~ x です．仮引数と引数をペアで指定する場合は，関数内の引数の順番を自由に変更できます．仮引数を指定しない場合は，help(glm) で表示される引数の順番を守って引数を並べなければなりません．

*40 これは正式には family = poisson(link = "log") とリンク関数も指定すべきなのですが，poisson family における default link function は "log" なので対数リンク関数を使いたいときは，とくにわざわざ指定する必要はありません．

*41 このオブジェクトの名前も何でもかまいません．また，fit に格納されている情報一覧を見るには，とりあえず names(fit) あるいは str(fit) としてみてください．

*42 この章の glm() の結果出力では，**逸脱度**（deviance） という「あてはまりの悪さ」の指標の表示を省略しています．逸脱度については，次の時間に説明しています．

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2917     0.3637    3.55  0.00038
x             0.0757     0.0356    2.13  0.03358
```

さて，この部分の読みかたを説明してみます[*43]．(Intercept) は切片 β_1 に，説明変数 x の係数は傾き β_2 に対応しています．

Estimate は 推定値のことで，結果出力をみると，最尤推定値は $\hat{\beta}_1 = 1.29$ と $\hat{\beta}_2 = 0.0757$ であるとわかります．

Std. Error はパラメーターの標準誤差の推定値です．**標準誤差**（standard error, SE）とは，この場合には推定値 $\hat{\beta}_1$ と $\hat{\beta}_2$ の「ばらつき」を標準偏差であらわしたものです．パラメーターの推定値のばらつきとは何でしょうか？ここでは，簡単に「同じ調査方法で同数の別データをとりなおしてみたりすると，最尤推定値もけっこう変わるので，そのばらつきぐあい」ということにしておきます．

この SE なる「パラメーターの推定値のばらつき」は，どのように推定されたのでしょうか．ひとつめの例題で説明したように，対数尤度は最尤推定値で最大値となる凸関数です．推定のばらつきが正規分布であると仮定し，さらに対数尤度関数は最大値付近でのカタチがその正規分布に近いと仮定すれば[*44]，上のような SE の推定値が得られます．

次にあらわれる z value は とよばれる統計量であり，最尤推定値を SE で除した値です．これによって，Wald 信頼区間というものを構成できます．図 16 で説明しますと，推定値たちがゼロから十分に離れているかどうかの粗い目安になります．この z 値は **Wald 統計量**（Wald statistics） ともよばれています．

最後の Pr(>|z|) は，この glm() の場合に限定して言えば[*45]，平均が z 値の絶対値であり標準偏差が 1 の正規分布における，マイナス無限大からゼロまでの値をとる確率です．この確率が大きいほど z 値がゼロに近くなり，推定値 $\hat{\beta}_1$ や $\hat{\beta}_2$ がゼロに近いことを表現するひとつの方法です．図 16 で説明すると，$\hat{\beta}_2$ の密度関数で黒くぬられた面積の 2 倍が，$\hat{\beta}_2$ に対応する z 値の Pr(>|z|) に相当します．

この確率 Pr(>|z|) をいわゆる P 値にみたてて，**統計学的な検定**（statistical test）[*46] ができると考える人もいます．しかんしながら，これはむしろ推定値の**信頼区間**（confidence interval） が近似的に算出されたぐらいに考えて[*47]，そのように結果を解釈するのが良いでしょう．

ある説明変数をモデルに含めるべきか否かといった判断は，このような Wald 信頼区間を使うのではなく，次の時間で説明するモデル選択を使ったほうが良いかもしれません．モデル選択はより良い**予測**をする統計モデルをさがしだそうとするもので,「この説明変数をいれるかどうか」といった判断はあてはまりの改善ではなく，

∗43　ここでは R の設定ファイル.Rprofile で options(show.signif.stars = FALSE) と指示しているので，こういう表の右端に「星」が表示されません．

∗44　このような仮定は正しいのでしょうか？たとえばサンプルサイズがそれほど大きくない場合は，このような Wald 統計量の方法は一種の近似と解釈したほうがよいでしょう．

∗45　glm() の family 指定で評価方法が変わります．たとえば， glm() で family = gaussian と指定した場合には，正規分布ではなく t 分布を使って確率を計算します．

∗46　統計学的な検定については次の時間に説明します．

∗47　パラメーターの値の最尤推定は点推定とよばれるのに対して，区間の推定は区間推定とよばれています．注意すべきは，α% 信頼区間とは，その区間内に「真の値がある確率が α%」という**意味ではない**ことです．信頼区間については統計学の教科書を参照してください．

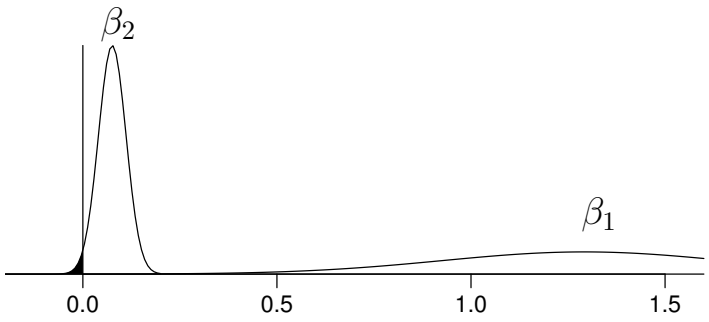

図 16　R の `glm()` のパラメーター推定値のばらつきの評価．たとえば，傾き β_2 の最尤推定値のばらつきが正規分布で近似できると仮定すると，確率密度関数（左）のようになる．この確率密度関数のスソの左側で黒く着色されている部分の面積の 2 倍が傾きの `Pr(>|z|)` に相当する（`Pr(>|z|)` $= 0.03358$）．確率密度関数（右）は切片 β_1 の推定値のばらつきに対応するものであり，`Pr(>|z|)` となる面積はほとんどゼロである（`Pr(>|z|)` $= 0.00038$）．

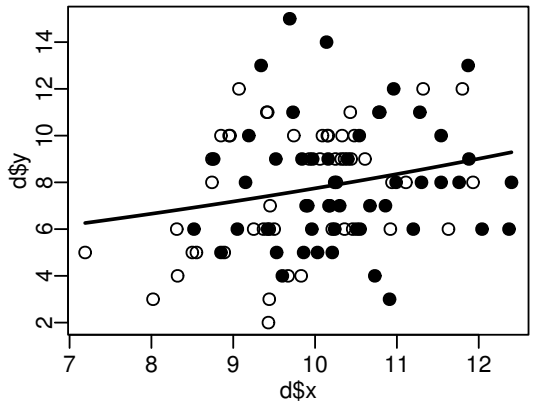

図 17　平均種子数 λ の予測．図 12 に λ の予測値（実線）を上がきしたもの．

予測の改善を目的としているからです．

この講義では，**最大対数尤度**（maximum log likelihood）を**あてはまりの良さ**（goodness of fit）と呼びます．あてはまりの良さが一番よくなるのは，対数尤度 $\log L(\beta_1, \beta_2)$ が最大になっているところであり，つまりパラメーターの値が最尤推定値 $\{\hat{\beta}_1, \hat{\beta}_2\}$ となっているときの対数尤度です．

R を使ってこのモデルの最大対数尤度を評価するには，

```
> logLik(fit)
'log Lik.' -235.3863 (df=2)
```

とすればよく，最大対数尤度は -235.4 ぐらいとわかります．`(df=2)` とは「**自由度**（degrees of freedom）が 2」をあらわしています．これは最尤推定したパラメーター数が 2 個（β_1 と β_2）である，ということです．

10.3 ポアソン回帰モデルによる予測

このポアソン回帰の推定結果を使って，さまざまなサイズ x における平均種子数 λ の**予測**（prediction）をしてみましょう．個体の体サイズ x の関数である平均種子数 λ の関数に推定値 $\{\hat{\beta}_1, \hat{\beta}_2\}$ を代入した関数

$$\lambda = \exp(1.29 + 0.0757x)$$

を使って R で図示してみましょう．

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> xx <- seq(min(d$x), max(d$x), length = 100)
> lines(xx, exp(1.29 + 0.0757 * xx), lwd = 2)
```

上のような操作によって，図 17 のように λ の予測値が曲線で示されます．あるいは，このように `predict()` 関数を使っても同じ結果が得られます．

```
> yy <- predict(fit, newdata = data.frame(x = xx), type = "response")
> lines(xx, yy, lwd = 2)
```

11 説明変数が因子型の統計モデル

次に，今まで放置していた施肥効果 f_i を説明変数としてくみこんだモデルも検討してみましょう．前に (p.16 で) 示したように，R 内で各個体の施肥処理は因子（factor）型のデータとして格納されています．もう一度確認すると，因子型変数の水準に `C` と `T` があり，`C` が 1 番目で `T` が 2 番目の水準という意味です．あとで推定結果を解釈するときには，このような因子型変数の構造を知っておかなければなりません．

このような因子型[*48]の説明変数は，GLM の中でどのようにあつかわれているのでしょうか．少し単純化した説明をすると，とくに指示をしない場合には，R の中で因子型の説明変数を含む線形予測子は，以下のような**ダミー変数**（dummy variable）におきかえられている — と理解しても大きな問題はありません[*49]．

あとから説明するように，実際の `glm()` 関数を使った推定計算では，ユーザーは何も考える必要はなく，データフレイム内の施肥処理の列 `f` を説明変数として指定するだけです．

しかし，簡単な場合の計算手順を明示的に説明するために，**あえて**この例題にそって因子型の説明変数をダミー変数におきかえてみましょう．植物の体サイズ x_i の効果を無視して，施肥効果 f_i だけが影響するモデルの平均値を

$$\lambda_i = \exp(\beta_1 + \beta_3 d_i)$$

と書くことにして，ここに登場する係数 β_1 は切片，β_3 は施肥の効果をあらわします．ここで説明変数が施肥処理 f_i ではなく，d_i はダミー変数におきかえられていて，以下のような値をとります:

*48 カテゴリ型変数とよばれることもあります．

*49 ただし R 内での実装はこの説明より複雑なものであり，対比（contrasts）という考えかたを使っています．ここで説明している水準間の比較方法は「最初の水準との比較」だけをあつかっていますが，これとは異なる水準間比較が必要になる場合があり，そのような状況に柔軟に対応するためです．

$$d_i = \begin{cases} 0 & (f_i = \texttt{C} \text{ の場合}) \\ 1 & (f_i = \texttt{T} \text{ の場合}) \end{cases}$$

いいかえると，個体 i が肥料なし（$f_i = \texttt{C}$）の場合は

$$\lambda_i = \exp(\beta_1)$$

となり，施肥処理した場合（$f_i = \texttt{T}$）の場合は

$$\lambda_i = \exp(\beta_1 + \beta_3)$$

となります．

話を R の glm() 関数を使った推定にもどします．先ほども述べたように，glm() 関数では，このような因子型の説明変数であっても，ダミー変数を準備するといった手間や工夫の必要もなく，モデル式を指定できます．

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

推定結果が格納されている fit.f の内容を出力してみましょう．

```
Call:  glm(formula = y ~ f, family = poisson, data = d)

Coefficients:
(Intercept)           fT
    2.05156      0.01277
...（以下略）...
```

パラメーターの推定値（Coefficients セクション）の出力をみると，施肥効果 f_i の係数の名前は fT となっていて，これは説明変数 f_i が T 水準でとる値を示しています．説明変数 f_i には C（肥料なし）と T（施肥処理）の 2 水準が設定されています．R は因子型説明変数 f_i の最初の水準 C の値をゼロとおき，これを基準にして T のような他の水準の値を推定します．もし個体 i の f_i が C ならば

$$\lambda_i = \exp(2.05 + 0) = \exp(2.05) = 7.77$$

であり，もし T ならば

$$\lambda_i = \exp(2.05 + 0.0128) = \exp(2.0628) = 7.87$$

となります．このように推定されたモデルでは「肥料をやると平均種子数がほんの少しだけ増える」と予測しています．

このモデルで最大対数尤度は

```
> logLik(fit.f)
'log Lik.' -237.6273 (df=2)
```

となり，(p.23 に示している) サイズ x_i だけモデルの最大対数尤度 -235.4 より小さく，あてはまりが悪くなっています．

この講義ではあつかいませんが，因子型説明変数の水準数が 3 以上になる場合もあります．たとえば，施肥処理する全個体に「肥料 A」を与えるのではなく，一部の個体にはそれとは異なる「肥料 B」を与えるとします．この場合，説明変数 f_i は $f_i \in \{\mathtt{C}, \mathtt{TA}, \mathtt{TB}\}$ といった具合に施肥処理について 3 水準が設定されます．

このような場合であっても，R のデフォルトの線形予測子のあつかいは，2 水準の場合を単純に拡張したものです．これも簡単のためダミー変数を使って説明すると，平均種子数は，

$$\lambda_i = \exp(\beta_1 + \beta_3 d_{i,A} + \beta_4 d_{i,B})$$

となり，係数 β_3 は肥料 A の効果，係数 β_4 は肥料 B の効果をあらわします．ダミー変数は以下のように設定されます．

$$d_{i,A} = \begin{cases} 0 & (f_i\text{が TA でない場合}) \\ 1 & (f_i\text{が TA の場合}) \end{cases}$$

$$d_{i,B} = \begin{cases} 0 & (f_i\text{が TB でない場合}) \\ 1 & (f_i\text{が TB の場合}) \end{cases}$$

このような 3 水準の因子型の説明変数を使った場合，R では `glm(y ~ f, ...)` と指定するだけで，β_3 と β_4 の推定結果が得られ，それぞれ `fTA` と `fTB` として表示されます．

12 説明変数が数量型 + 因子型の統計モデル

今度は，個体の体サイズ x_i と施肥効果 f_i の複数の説明変数を同時にくみこんだ統計モデルを作ってみましょう[*50]．

GLM では，複数の説明変数の効果は線形予測子の中で和として表現します．この例題の場合，

$$\log \lambda_i = \beta_1 + \beta_2 x_i + \beta_3 d_i$$

となり，β_1 が切片に該当する部分，β_2 がサイズ（x_i）の効果で，β_3 が施肥処理 (2 水準の f_i をダミー変数化した d_i) の効果となります[*51]．

R の `glm()` 関数による推定計算は特に何も指示しないで，モデル式の部分を `x + f` とするだけで適切に処理してくれます．

```
> fit.all <- glm(y ~ x + f, data = d, family = poisson)
```

結果を格納している `fit.all` の出力を見てみましょう．

*50 複数の説明変数をもつ統計モデルによるあてはめは**重回帰**（multiple regression） とよばれることがあります．この講義では説明変数の個数に関係なく回帰とよびます．

*51 サイズと肥料の交互作用項については，この章では説明しません．

```
Call:  glm(formula = y ~ x + f, family = poisson, data = d)

Coefficients:
(Intercept)          x          fT
    1.2631      0.0801     -0.0320

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:      89.51
Residual Deviance: 84.81        AIC: 476.6
```

この結果出力をみると，前の節では肥料の効果 fT がプラスであったのに，このモデルではマイナスだと推定されています．肥料の効果についてはいよいよわからなくなりました．

このモデルで最大対数尤度は，

```
> logLik(fit.all)
'log Lik.' -235.2937 (df=3)
```

となり，p.23 に示している x_i だけのモデルの対数尤度 (-235.4) と比較すると，少しだけあてはまりが良くなっています．モデルごとに異なるあてはまりの良さの比較については，次の時間に検討します．

12.1 対数リンク関数のわかりやすさ: かけ算される効果

この章の 10 節で,「 対数リンク関数では要因の効果が積であらわされる」と述べました．この数量型 + 因子型モデルの推定結果を使って，それを説明してみましょう．

推定計算の関数 glm() でモデル式を glm(y ~ x + f, ...) と指定しているので，説明変数の効果を足し算であらわしているように見えます．しかし， 対数リンク関数を使っているので，足し算ではなくかけ算で要因が平均に効果を及ぼしています．

このモデルの推定結果を予測としてまとめると，体サイズ x_i の個体 i の施肥処理 f_i が C ならば平均種子数は，

$$\lambda_i = \exp(1.26 + 0.08x_i)$$

であり，もし T ならば，

$$\lambda_i = \exp(1.26 + 0.08x_i - 0.032)$$

となり，以下のように分解できます．

$$\begin{aligned} \lambda_i &= \exp(1.26) \times \exp(0.08x_i) \times \exp(-0.032) \\ &= \text{（定数）} \times \text{（サイズの効果）} \times \text{（施肥処理の効果）} \end{aligned}$$

サイズ x_i が増加する影響は「説明変数 x_i が 1 増加すると，λ_i は $\exp(0.08 \times 1) = 1.08$ 倍に増える」と予測されますから，説明変数の増分は足し算ではなくかけ算のかたちで平均を変えています．施肥処理の効果も足

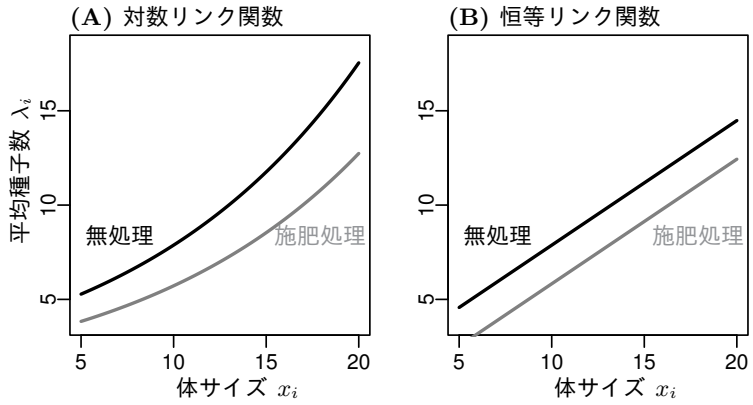

図 18　リンク関数がちがうと予測内容が変わる．(A) 対数リンク関数 (B) 恒等リンク関数（リンク関数なし）それぞれの予測．図示をわかりやすくするために施肥処理の効果を 3 倍に設定，つまり肥料の悪影響が 3 倍になっている．

し算できいているのではなく，かけ算で影響しています．この場合ですと $\exp(-0.032) = 0.969$ ですから，肥料をやると種子数の平均が 0.969 倍になると予測されます．

平均 λ_i はサイズ・施肥処理それぞれの効果の積ですから，図 18 (A) のようになります[*52]．サイズ x_i が大きいほど施肥効果あり・なし間の乖離が大きくなります．

もし対数リンク関数を使わなかったらどうなるのでしょうか？このように平均が線形予測子に等しい，つまりリンク関数がとくに**何もない**場合，R ではこの状態を 恒等リンク関数（identity link function）とよびます．この **恒等リンク関数** を使ってポアソン回帰をしてみましょう[*53]．すると平均種子数 λ_i の予測は，$\lambda_i = 1.27 + 0.661x_i - 0.205d_i$ であり，図 18 (B) のようになります[*54]．

この恒等リンク関数を使ったモデルの主張は，無処理における平均種子数が 0.1 個だろうが 1000 個だろうが，施肥処理をやったらどちらも 0.205 個減って -0.1 個とか 999.8 個になるんだ— ということです．複数の効果がかけ算で影響すると考えている，対数リンク関数のポアソン回帰とはまったく相違するモデルです．

どちらのリンク関数が「妥当なモデル」なのかは，あてはまりの良しあしだけで決まる問題ではありません．上の比較から考えると，対数リンク関数のほうが「まし」な統計モデルのような気もします．いずれにせよ重要なのは，パラメーター推定（あるいは「検定」）ができればどんなモデルでもいいという発想はやめて，数式が現象をどのように表現しているのかという点に注意しながら統計モデルを設計することです．

13 「何でも正規分布」「何でも直線」には無理がある

GLM において，確率分布は等分散の正規分布かつ「リンク関数なし」— つまり 12.1 項で登場した 恒等リンク関数 と指定すると，これは一般化（**generalized**）ではない**線形モデル**（linear model, LM）あるいは一

*52　図のキャプションに書いているように，リンク関数の効果を見えやすくするために，施肥処理の効果を 3 倍（-0.032×3）にしています — 肥料の「有害性」が 3 倍だと考えてください．

*53　`glm()` 関数で `family = poisson(link = "identity")` と指定します．この例題のデータセットでは問題ありませんが，データによっては λ が負の値となり推定計算ができません．

*54　(A) にあわせて施肥効果の処理を 3 倍，つまり -0.205×3 としています．

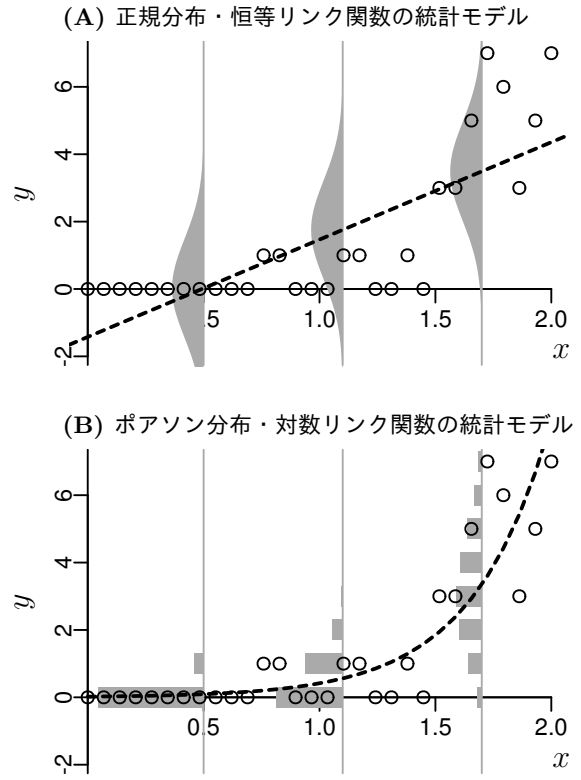

図 19　回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで示しているのは $x \in \{0.5, 1.1, 1.7\}$ での y の確率分布または確率密度関数．(A) データのばらつきが等分散正規分布，y の平均が $\beta_1 + \beta_2 x$ であると仮定したモデルのあてはめ (B) データのばらつきがポアソン分布，y の平均が $\exp(\beta_1 + \beta_2 x)$ であると仮定したモデルのあてはめ．

般線形モデル（**general** lienar model）とよばれます．ここでは，とくによく使われる LM のあてはめのひとつである**直線回帰**（linear regression） ポアソン回帰を比較してみましょう．

この章の例題とはまた別の架空データがあるとしましょう．図 19 に示しているデータ点（図中の丸）(x_i, y_i) があたえられたときに，図 19 (A) のように「とにかく散布図に直線をひけばいい」という発想で直線回帰をしてしまう人をよく見かけます．その正当化の理由として,「世の中は何でも正規分布だから」などという人もいます．このような「直線さえひければ何でもいい」という方針の人たちの流儀は正しいのでしょうか？

直線回帰は GLM の一部なので[*55]，そこで使われている統計モデルの特徴を列挙してみましょう:

- 観測値 $\{x_1, x_2, \cdots, x_n\}$ と $\{y_1, y_2, \cdots, y_n\}$ のペアがあり，$\mathbf{X} = \{x_i\}$ を説明変数，$\mathbf{Y} = \{y_i\}$ を応答変数とよぶ
- $\mathbf{Y}$ は平均 μ_i で標準偏差 σ の正規分布にしたがうと仮定する
- あるデータ点 i において平均値が $\mu_i = \beta_1 + \beta_2 x_i$ となる

このように整理すると直線回帰に使われる統計モデル LM が GLM の一部であることがよくわかります[*56]．

---

*55　**??** 節も参照．

*56　LM を使ったデータ解析について少し補足します．複数の数量型の説明変数がある場合は**重回帰**（multiple regression）．説明変数 x_i が因子型であるモデルのあてはめは**分散分析**（ANOVA）．また説明変数がふたつ以上あり，かつ数量型・因子型の説明変数が混在

このように統計モデルで仮定していることが明らかになれば,「どんなデータでも直線回帰」という作法の限界が見えてきます．たとえば，データの図 19 に示されている，応答変数 y が 0 個, 1 個, 2 個と数えられるカウントデータであるとしましょう．そのようなデータに対して,「何でも正規分布」「x と y はいつも直線関係」と仮定するのは無理があり，実際に直線回帰による y の平均値の予測（図 19 A の破線）を見ると，以下のように，いろいろとおかしな点を指摘できます:

- 正規分布は連続的な値をあつかう確率分布だったはずでは？
- カウントデータなのに平均値の予測がマイナスになる理由は？
- 図でみると「ばらつき一定」ではなさそうなのに，分散一定を仮定？

つまり，図 19 に示されているデータを表現する手段として，直線回帰の統計モデルは「現実ばなれ」しています[*57].

直線回帰をすればパラメーターの推定値を得られますが，そもそも「現実ばなれ」した統計モデルを使っているので，解析そのものに意味がないということです[*58].

これに対して，図 19 (B) で示しているように，このデータをポアソン分布を使った GLM で説明しようとするのは比較的妥当なものであり，なぜならば上にあげた問題点は,

- ポアソン分布を使っているのでカウントデータに正しく対応
- 対数リンク関数を使えば平均値はつねに非負
- y のばらつきは平均とともに増大する

このようにうまく解決できているように見えるからです.（図 **??** における最初のステップアップ）.

また，応答変数 y を $\log y$ のように**変数変換**して直線回帰することと，ポアソン回帰はまったく別ものであることに注意してください．試してみればわかりますが，両者の推定結果は一致しません．とくに y がゼロに近い値での対数変換はデータのばらつきの「見た目」をすこし変えるだけのものであり[*59]，わざわざ無理矢理に正規分布モデルをあてはめる利点がありません．このような強引な変数変換わざを避け，y の構造にあわせて適切な確率分布を選ぶというのが，この講義で強調している統計モデリングの方針です.

この章で紹介した GLM の特徴は，データにあわせて確率分布とリンク関数を選べる点にあり,「何でも正規分布」「とにかくセンをひけばいい」という発想から脱却する最初の一歩となるでしょう．ポアソン分布以外の確率分布を部品とする GLM については，明日紹介します.

しているモデルなら**共分散分析**（ANCOVA）.

*57　R の `glm()` を使って直線回帰をする場合には，`family = gaussian` と指定します．また `family = gaussian(link = "log")` とすれば 対数リンク関数を指定できます．これで平均値がマイナスになる問題は回避できますが，他の問題は依然として解決しません.

*58　このような現実ばなれした直線回帰で得られる R^2 値だの P 値だのも同様にナンセンスです.

*59　そもそも $\log 0 = -\infty$ となりますが.