本書の構成

最初から、最後まで。これが本書の方針です。
統計学の基礎から始めて、一般化線形モデルまで、順を追って解説します。

「第 1 部　統計学の基礎と検定の考え方」において、データ解析の基礎を解説します。統計モデルに移る前段階となります。
平均値や期待値、分散といった用語から始まり、「検定」の考え方、「R」という統計ソフトの使い方まで解説します。

「第 2 部　統計モデル基礎：正規線形モデル」「第 3 部　正規線形モデルによるデータ解析」では、分散分析や回帰分析といった統計モデルの基礎から始め、統計モデルを用いたデータ解析の応用までを学びます。

第 4 部、5 部は、確率分布を中心とした統計学基礎を学ぶパートです。
「第 4 部　確率と統計データ」では、実務的なデータ解析と直接は関係のない統計解析の基礎理論を、「第 5 部　確率分布と統計モデル」では、基礎理論を実際のデータ解析に結び付ける方法を学びます。

第 6 部、第 7 部は、一般化線形モデルのパートです。
「第 6 部　一般化線形モデル」で一般化線形モデルの基礎を、「第 7 部　一般化線形モデルによるデータ解析」において、その応用を学びます。

最後の「第 8 部　情報理論と統計学」では、さらに進んだ話題として、赤池の情報量規準（AIC）の考え方とその使用法を説明します。

本書では「R」と呼ばれる無料の統計解析ソフトを使います。本書を読めば、一般化線形モデルをパソコンで計算する技術が身につくでしょう。
本書で扱った「R」のソースコードやデータに関しては、著者のウェブサイト（http://logics-of-blue.com/）からダウンロードすることができます。また、巻末には「逆引き R 関数」を載せました。合わせてご参照ください。

第４部　確率と統計データ

２章　データが得られるプロセス

この章ではデータが手に入るプロセスを理解していただきます。

あなたが手に入れたデータは、どのような経緯であなたの手元にやってきたのか、その流れを理解してください。

2ー1　　母集団と標本とサンプリングの関係

データが手に入るプロセスを理解するためには、母集団と標本、そして確率分布と確率変数の関係を理解する必要があります。
順番に行きましょう。

データが手に入るプロセスを理解する最初のステップが母集団です。

母集団とは「すべてのデータ」のことです。
例えば湖の中にいるすべての魚を母集団としましょう。

湖で釣りをします。5 尾釣れました。この 5 尾は標本と呼ばれ、釣りをしてデータを得ることをサンプリングと呼びます。
今回は 5 尾釣ったので、サンプルサイズは 5 となります。

湖の中のすべての魚（母集団）からサンプリングをして、データ（標本）を手に入れました。
これがデータの取得という行為を理解するための最初のステップです。

2ー2　　母集団からのサンプリングがイメージしにくい例

しかし「母集団から標本をサンプリングする」という状況が、想像しづらい時

があります。

サイコロの例を挙げます。
サイコロを投げて、どのような目が出たのかを確認するという行為を想像してください。

ここでの標本が何なのかを想像することは簡単です。「サイコロを投げて出た目」が標本です。6 の目が出たのであるならば、6 の目という標本がサンプリングされたことになります。

では、母集団とは何でしょうか？
湖の例との違いを確認してください。
湖では「湖の中にいるすべての魚」が母集団でした。イメージすることは容易です。
しかし、サイコロの場合は事情が違います。
サイコロは何億回、何兆回、何京回でも投げることができます。サイコロの出る目の母集団は無限個あります。
では、「サイコロを投げて出た目」の母集団とは何でしょうか？

この場合の母集団は、この世界に存在するすべての「サイコロを投げたという試行」です。それは 2000 年 1 月 1 日午後 5 時 56 分に東京で投げたものかもしれないし、2100 年 4 月 24 日午後 3 時 6 分にリオデジャネイロで投げられるものかもしれません。
これらすべての「サイコロを投げたという試行」を想定し、そこから一部のデータを抽出して標本とする。それがサイコロのサンプリングです。

……何を言っているのかわからないと感じて当然です。
母集団という存在は大変にわかりづらいです。大変に扱いづらいです。
もっと簡単に、母集団という存在を取り扱う方法はないのでしょうか。

2−3　　無限母集団と確率分布と確率変数の関係

母集団からのサンプリングという概念は、確率分布を使うことにより、スマートに表すことができます。

サイコロのサンプリングとは以下の確率分布から確率変数を得ることです。

$$p_i = \left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$$

確率変数は以下の値をとります。

$$x_i = \{1,2,3,4,5,6\}$$

サイコロを無限回投げたとします。その「無限回サイコロを投げた結果」が母集団です。

母集団において（無限個あるので厳密には数えられないけれど）数えてみたら1/6 の割合で 1 の目が存在することがわかりました。

であるならば、「母集団からランダムにデータを抽出する」という行為は、「1/6 の確率で 1 の目が出てくる」状況と同じです。

そこで、「データが得られる確率分布」を母集団の代わりに使用するというアイディアが生まれました。

このアイディアを用いると、扱いづらい母集団という存在を直接用いる必要がなくなります。（中略）

3章　データを解析するプロセス

データが得られるプロセスを理解できれば、データを解析するプロセスを理解することは容易です。

データ解析とは、データが得られるそのプロセスを逆算して求めることです。

手持ちのデータ（標本）からデータが発生されたプロセス（母集団、あるいは母集団分布）を推定することこそが、推測統計学におけるデータ解析の目的です。

3−1　　手に入れていないデータを扱う方法

データ解析の本当の対象は「あなたが手に入れていないデータ」です。

しかし、「あなたが手に入れたデータ」から直接「手に入れていないデータ」を推定することには無理があります。

それは例えば、いま東京で投げられたサイコロの目をよく眺めても、2050 年 5 月 12 日 15 時 24 分にヨハネスブルグで投げられるサイコロの目が推定できないことからも、十分明らかです。

それでも、「手に入れていないデータ」に近づく手段があります。

そのカギとなるものが、母集団の確率分布です。

3−2　　データから母集団分布を推定する方法

サイコロを無限回投げることはできません。でも、600 回くらいなら投げられます。

その時、100 回 1 の目が出たとします。

すなわち、1 の目が出た確率は 6 分の 1 です。

また、2 の目も 3 の目も、6 分の 1 出ました。

良く数えてみると、1~6 の目は全て「6 分の 1」という確率で現れていたとします。

そうしたら、この「手持ちのデータから計算された確率分布」を「母集団の確率分布」の代わりに使えるのではないか。

この発想が、推測統計学です。（中略）

第６部　一般化線形モデル

1章　一般化線形モデルの長所

一般化線形モデルの詳細を学ぶ前に「なぜ一般化線形モデルが必要とされるのか」を説明します。

1−1　　一般化線形モデルとは

一般化線形モデルとは、母集団の確率分布に正規分布以外の確率分布を用いることができる線形の統計モデルのことです。

今まで正規分布だけに絞っていたのが、他の分布にも使えるようになったので「一般化」されたわけです。

1−2　　一般化線形モデルの使い時

正規分布以外の確率分布を使う理由を知るためには「正規分布を使うべきでない理由」を理解するのが手早いです。

正規分布は連続データを対象としていましたが、例えば猫が 0.25 匹いるという状況は想定しづらいです。また、正規分布のとりうる範囲は−∞〜∞まででした。しかし、猫の数が−3 匹では困ります。

他にも、「コインを投げて表が出る確率が−324%」とった状態も正規分布だとあり得てしまいます。

このように正規分布では「本来とってはいけない値」をとってしまうことがたびたびあります。

こんな時には「小数点以下の値をとらない確率分布」などを使用できれば便利です。

という訳で、データに合わせて「変な値が出ないようにモデルを柔軟に組換えよう」というニーズから一般化線形モデルが使われるということです。（中略）

3章　一般化線形モデルの推定

3−1　　ポアソン分布

母集団が正規分布以外の場合でも、最尤法を用いればパラメタの推定ができます。具体的にどのような確率分布を相手にするかを説明します。

本書で最初に扱うのはポアソン分布です。正規分布では扱いにくいデータを対象とするのに向いています。

ポアソン分布は、個体数や回数を表す確率分布です。

猫が 2 匹居るとか、商品が 4 個売れた、1 時間にウェブページが 5 回表示されたといった「●個」「●回」単位のデータを対象とします。

正規分布は連続データを対象としていましたが、猫が 0.25 匹いるという状況は想定しづらいです。また、正規分布のとりうる範囲は−∞～∞までででした。しかし、猫の数が−3 匹では困ります。

そこで、ポアソン分布を用います。

ポアソン分布は離散変数を対象とする確率分布です。離散変数とは「1 個」「2 個」ととびとびのデータを指します。1.5 個という中間が存在しない変数です。そのため、小数点以下のデータは発生しません。

また、ポアソン分布は常に値が 0 以上であることを想定しています。

なので、個体数や回数を表すのに大変便利です。

ポアソン分布の確率密度関数は以下の通りです。ただし x は確率変数です。

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

推定すべきパラメタはλのみです。

なお、ポアソン分布においては、期待値も分散もともにλと等しくなります。

また、ポアソン分布に従う確率変数は 3 個体とか、6 回といった正の整数しかとりません。連続変数ではないので確率を計算するのに面積を求める必要はありません。x に 3 を代入することで、「猫が 3 個体いる確率」を直接計算することが

できます。

一般化線形モデルではこのほかに「二項分布」なども使用できます。
二項分布はコインの裏表や病気にかかったかかかっていないか、など 2 択で表すことのできるデータに対して適用されます。本書後半で使用例を載せます。

次からは母集団の確率分布にポアソン分布を仮定した一般化線形モデル（ポアソン回帰）を例として説明します。
しかし、一般化線形モデルはポアソン分布以外にも（本書で紹介している以外でも）いろいろな確率分布が使えることだけ、頭の片隅に入れておいてください。

3−2　線形予測子

統計モデルとは確率分布の候補です。
統計モデルにおいて推定すべきは確率分布のパラメタです。

ここで「確率分布のパラメタを変化させたい」という要望が生まれたとします。
例えば「気温の高低により、ビールの売り上げの期待値が増減する」という関係を表したかったとします。
この要望を満たす一つの方法が、線形予測子の使用です。

線形予測子は例えば以下のような構造をとります。

$$\text{ビールの売り上げの期待値} = a \times \text{気温} + b$$

母集団の確率分布に正規分布を仮定したとしましょう。すると、気温が 20 度の時におけるビールの売り上げの期待値は「$20a+b$」となります。
確率分布のパラメタを推定するために、傾きや切片を推定しているということです。

これは母集団の確率分布が変わっても同じです。
母集団の確率分布にポアソン分布を仮定したとしましょう。
ポアソン分布において推定すべきは期待値λです。「餌の多少により、猫の個体数が増減する」という関係を表すには、以下のような線形予測子を使用します。

$$\text{ノラ猫個体数の期待値} = a \times \text{餌の量} + b$$

餌の量が 5 だった場合には、ノラ猫個体数は「期待値 λ が $5a+b$ のポアソン分布」に従います。

分散分析モデルで出てきたような選択肢を入れることも可能です。

$$\text{ノラ猫個体数の期待値} = a \times \text{餌の量} + b \times \text{草地} + c$$

この場合は、場所が草地なら 1、それ以外の場所なら 0 を代入してやればよいです。草地以外の場所と比べて、草地だとどれだけ猫が増えるのかがわかります。

統計モデルによる予測値は、確率分布のパラメタである点を思い出して下さい。そうすれば「線形予測子」という言葉もしっくりくると思います。

なお、ポアソン分布を仮定した場合でも、それ以外の確率分布を使用した場合でも、この線形予測子の構造はすべて共通です。新しいことを覚える必要はありません。

この使い勝手の良さも、一般化線形モデルが支持される理由の一つでしょう。

3−3　　リンク関数

線形予測子でパラメタを変化させるのはよいのですが、例えばポアソン分布の場合だと、期待値が負になると困ります。猫の個体数の期待値が−3 というのは、好ましくありません。

そこで、リンク関数をかませることによって「確率分布のパラメタが取りうる範囲」を制限します。

ポアソン分布の場合は、期待値が常に正でなくてはならないという制約があります。この制約を満たすためには指数を使うのが便利です。

$$\text{ノラ猫個体数の期待値} = \exp(a \times \text{餌の量} + b)$$

なお、exp（×）で e の×乗を表します。R の記法に合わせました。

e は大体 2.7 であり、符号はプラスなので、e の×乗は常にプラスになります。これならば、猫の個体数の期待値がマイナスにならなくて済みます。

しかし、このままだと線形予測子が何なのかわかりづらいので、以下のように書き直すことが普通です。

$$\ln(\text{ノラ猫個体数の期待値}) = a \times \text{餌の量} + b$$

この時「リンク関数がログである」と呼びます（ln は底が e である対数です）。
リンク関数をログにすると、期待値が負になる状況を防ぐことができます。

リンク関数にはログ以外にもいろいろな種類があります。
別の例は後程ロジスティック回帰の章で説明します。（中略）

3−7　R によるポアソン回帰 1：計算の流れ

最尤法を用いた一般化線形モデルの推定手順は、最小二乗法を用いた回帰分析の推定手順とたいへんよく似ています。
下記の順に解析を行います。

1. データの準備（プロットは省略）
2. 対数尤度の計算
3. 対数尤度計算関数の作成
4. optim 関数を用いたパラメタ推定

3−8　R によるポアソン回帰 2：データの準備

今回も架空の観測データを使います。
敷地内の餌が増えることにより、そこに住む猫の数が増えるかどうかを調べたい、というテーマです。

以下のようなデータを使用します。

```
d6 <- data.frame(
  esa = c(1, 2, 3, 4),
  neko = c(4, 10, 7, 14)
)
```

3−9　　R によるポアソン回帰 3：対数尤度の計算

「ノラ猫個体数の期待値 λ 」が 5 の時、今回の猫の個体数データが得られる確率を各々求めてみます。

```
> dpois(lambda=5, d6$neko)
[1] 0.1754673698 0.0181327887 0.1044448630 0.0004717363
```

続いて、尤度を計算します。中身を掛け合わせればよいです。

```
> pPois <- dpois(lambda=5, d6$neko)
> likelihood <- pPois[1] * pPois[2] * pPois[3] * pPois[4]
> likelihood
[1] 1.567644e-07
```

期待値が 5 のポアソン分布を仮定すると、0.00000016 ほどの確率で、今回のデータが生じるという結果になりました。

次は対数尤度の計算です。対数をとるだけです。

```
> log(likelihood)
[1] -15.66852
```

これは以下の計算結果と一致します。

```
> sum(log(dpois(lambda=5, d6$neko)))
[1] -15.66852
```

log 関数で対数をとってから sum 関数で合計しました。対数をとると掛け算が足し算に変わるというところだけ注意してください。

次はλを以下の線形予測子に従って変化させたとしましょう。

$$0.3 \times \text{餌の量} + 1$$

この時の予測値（ノラ猫個体数の期待値：λ）は以下のように計算されます。

```
> a <- 0.3
> b <- 1
> lambdaHat <- exp(a*d6$esa + b)
```

```
> lambdaHat
[1] 3.669297 4.953032 6.685894 9.025013
```

リンク関数がログの場合は、線形予測子に exp を取ることに注意してください。

餌の量のデータは『esa = c(1, 2, 3, 4)』でした。

先ほどの計算により、以下の 4 つの予測値が求まったということです。

・餌の量が 1 の時の猫の個体数の期待値 λ ：約 3.7
・餌の量が 2 の時の猫の個体数の期待値 λ ：約 5.0
・餌の量が 3 の時の猫の個体数の期待値 λ ：約 6.7
・餌の量が 4 の時の猫の個体数の期待値 λ ：約 9.0

この時の「データが得られる確率」、すなわち尤度は以下のようにして計算されます。

```
> dpois(lambda=lambdaHat, d6$neko)
[1] 0.19255942 0.01729315 0.14792413 0.03283584
```

猫の個体数データは『neko = c(4, 10, 7, 14)』でした。

先ほどの計算により以下の 4 つの確率が求まったということです。

・λが約 3.7 の時に猫の個体数が 4 になる確率 ：約 0.19
・λが約 5.0 の時に猫の個体数が 10 になる確率 ：約 0.02
・λが約 6.7 の時に猫の個体数が 7 になる確率 ：約 0.15
・λが約 9.0 の時に猫の個体数が 14 になる確率 ：約 0.03

この 4 つの確率を使って対数尤度を求めます。

```
> sum(log(dpois(lambda=lambdaHat, d6$neko)))
[1] -11.03209
```

1. 予測値（ノラ猫個体数の期待値：λ）を計算する
2. 予測された λ を用いて「データが得られる確率（尤度、あるいは対数尤度）」を計算する

この流れを覚えておいてください。

3−10　R によるポアソン回帰 4：関数の作成

対数尤度の−1 倍を計算する関数を自作します。

マイナス 1 倍しているのは、後で optim 関数を使う都合です。optim 関数は「ある値を最小にするパラメタ」を推定するものです。「最大にするパラメタ」を推定したいのであれば、マイナス 1 をかけて最大最小を逆にする必要があります。

```
1  calcLogLikPoisson <- function(para){
2    a <- para[1]
3    b <- para[2]
4    lambdaHat <- exp(a*d6$esa + b)
5    logLik <- -1*sum(log(dpois(lambda=lambdaHat, d6$neko)))
6    return(logLik)
7  }
```

1 行目：関数を作るという指定

2〜4 行目：予測値の作成

5〜6 行目：対数尤度に−1 をかけたものを計算して結果を返す

3−11　R によるポアソン回帰 5：パラメタ推定

最後に、optim 関数を使って、対数尤度を最大にする係数を求めます。

```
> optim(c(0.3, 1), calcLogLikPoisson)
$par
[1] 0.3172607 1.3138120
$value
[1] 8.970149
$counts
function gradient
     59       NA
$convergence
[1] 0
$message
```

```
NULL
```

傾きが大体 0.3 で切片が 1.3 となりました。

対数尤度は－9.0 くらいです。

3－12　R の glm 関数によるポアソン回帰

次は R にもともと用意されている glm 関数を用いて、ポアソン回帰モデルを推定してみます。

```
modelPoisson <- glm(neko ~ esa, family="poisson", data=d6)
```

正規線形モデルと異なるのは以下の 2 点です。

1. 使用する関数が glm 関数に変わった
2. 確率分布を family で指定した

なお、リンク関数については、確率分布をポアソン分布に変えた時点で自動的にログが適用されます。

推定結果はこちら。

```
> modelPoisson
Call: glm(formula = neko ~ esa, family = "poisson", data = d6)
Coefficients:
(Intercept)         esa
      1.3138      0.3173
  Degrees of Freedom: 3 Total (i.e. Null);  2 Residual
  Null Deviance:      6.445
  Residual Deviance: 2.221        AIC: 21.94
```

下の方に出てきた Deviance などの指標については次の章以降で説明します。まずは推定された係数が optim の結果と同じになっていることを確認してください。（以下略）

出版社 Web サイト　：http://www.pleiades-publishing.co.jp/

書籍のサポートサイト：http://logics-of-blue.com/平均・分散から始める一般化線形モデル入門：サ/