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 2 回 (2008–10–30)

さまざまな確率分布と最尤推定

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 2 回目です．R のインストールはうまくできましたか? 今日はこの R を動かしながらハナシをススめていきたいと思います．

前回は,

- データ解析とは，観測データをうまく説明できるような統計モデルの構築 (統計モデリング) にほかならない
- 統計モデルの基本部品は**確率分布**である

といったことを述べてみました．今日はその**基本部品**たる確率分布とやらを説明してみたいと考えています．

確率分布，さてさてどう説明したものでしょうか……うーむ，ハナシをたいへん**乱暴**にススめて恐縮なんですけれど,「確率とは何か? 確率分布の厳

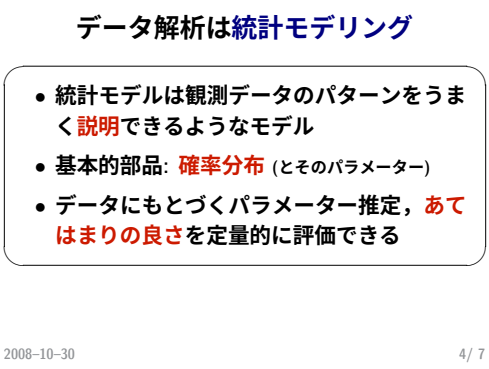


図 1: 統計モデルとは何か?

密な定義は?」ははぶいてしまいましょう……というのも，ホントにゼロに近いところから出発して話していくと，おそらく聴かされてるほうは面倒で退屈でわかりにくい，と思うわけですよ．[1]

こういうふうに乱暴に始めるのはいいんだけど，どういう確率分布をもちだすのが説明に都合よろしいでしょうかねえ……それではポアソン分布の説明から始めてみましょうか．前回も述べましたが，この講義では **ポアソン分布・二項分布・正規分布**という三種類の確率分布だけをあつかうことにしよう，と計画しています．[2]

1. とはいえ，このあたり，前回示したような統計学教科書などあとからながめてみて，自分でおぎなえるところはおぎなってみてください．

2. 今回は全 7 回から全 5 回になったので，正規分布を使った統計モデリングはほとんど登場しません．

1. いきなり R & ポアソン分布で統計モデリング?!

いきなり統計ソフトウェア R なんかを起動してそこに何かデータが入っていたとしましょう[3] ……そうですね，私が誰かから R のデータ[4] をもらってしまって，それは「data と名づけられた[5] vector object」になっていてこれは「ある人が毎日うけとったメイル (e-mail) の数，50 日ぶん」である，と．何だかよくわかりませんが，とりあえず R でそのデータを開いてみましょう．[6]

```
> 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
```

これで意味不明ぎみだった「data と名づけられた vector object」なるものの正体が少しわかったと思います．これは 0 個，1 個, 2 個，… などと数えられるデータ，つまり **カウントデータ** (ここでは一日に受けとったメイルの本数) になっていて，それが vector なる一次元の構造をもつデータ構造

3. さすがにわれながら不条理な状況設定だと認識しております．

4. これはホントの観測データではなく，講義用の架空データです．どうやって作るかは後述．

5. データを入力したヒトが勝手につけた名前，と考えてください

6. 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．

に格納されていたらしい，その vector object は data と名づけられている，という状況におかれているわけです．

上の R のユーザーインターフェイス上で示されていることは，ユーザーが data と R のコマンドプロンプト[7] 上で入力すると[8] ,「data の内容はこうです」が示されています．たしかに 50 個の整数からなっているように見えますね．

この data をデータ解析しなければならぬ状況におかれている，とします．さっそく R を使って調べてみましょう．[9]

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

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

length() 関数によってこの data に含まれるデータ数は 50 個だと確認ができ，summary() 関数によって，平均値や最小値・最大値・四分位数などがわかります．

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．[10] Median は中央値・中位値と呼ばれる推定値ですね．そして (標本) **平均値** (Mean) が 3.56 である，と．ついでながら (標本) **分散** (variance) はこのように計算できます．

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

この標本分散の推定値は data の「ばらつき具合」を表現しています．なお (標本) **標準偏差** (standard deviation; SD) とは分散の平方根です．R ではこんなふうに計算できます．

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

ここで重要なのは平均値なんかだけをながめてわかったような気分になら

7. R 上の > に命令を入力すると R がそれを実行してくれる，というインターフェイスのことです．

8. 自動的に print(data) が実行されています．

9. R では # のうしろから行末までに書かれてることは読み飛ばされます．# はコメントマークと呼ばれています．

10. この値は観測値そのものではなく，たいていの場合，補間された値になっています．

ないことです．解析しなければならないデータはもっと**「見えるカタチ」**にしましょう．たとえば「5 通うけとったのは 50 日中の何日だったのか?」といったことも調べる必要があります．このような度数分布を R で調べる方法はたくさんあるのですが……たとえば一番簡単には，[11]

```
> summary(as.factor(data))
 0  1  2  3  4  5  6  7
 1  3 11 12 10  5  4  4
```

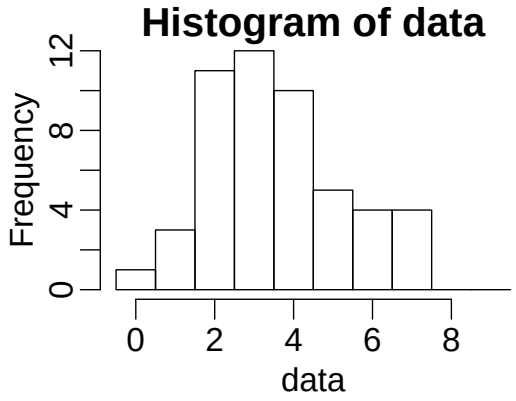
とすればよいだけです．あるいはもっと簡単に

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

としてもいいです．

前回に皆さんに「実践してもらいたいこと」としてあげた**「観測データの図をたくさん作ろう」**にしたがって，この data を度数分布図 (histogram) として図示してみましょう．[12]

```
> hist(data, breaks = seq(-0.5, 8.5, 1))
```



R はホントに便利ですね．

ここまででわかったことは，某氏が毎日うけとるメイルの本数 (データは 50 日ぶん) は

- 一日にうけとるメイル数の平均値は 3.56 通
- しかしそれより多い日も少ない日もある

ということで，何か確率論的な (probablistic) モデルでこの現象を説明できそうだ，つまり統計モデリング可能だろう，とハナシをつなげたいわけです．

11. summary() 関数の挙動が先ほどと異なってますよね……「どうしてこうなってしまうんだろう」と苦悩してくださいね :-)

12. seq(···) の意味がわからないヒトは自分の R で seq(-0.5, 8.5, 1) を入力して実行してみてください．

この観測されたパターンはどういった確率分布によって表現できるのでしょうか?

こういうカウントデータの統計モデリング，その基本部品となる確率分布としては (とりあえず) **ポアソン分布** (Poisson distribution) を使ってみる [13] というのがよくある統計モデリングの手口です．いささか強引ですが，ここでは「そういうものだ」というコトにしてハナシを進めていくことにします．

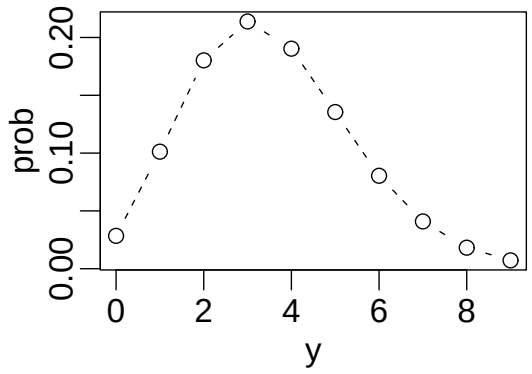
ポアソン分布っていうのはいったいぜんたい何なのか? いきなり図でも示して説明してみましょうか.

たとえば，平均 3.56 のポアソン分布 [14] とはこういうものです.

13. すでにお忘れかもしれませんが,「世の中にはポアソン分布と呼ばれる確率分布がある」なるとうとつなる前提でハナシが展開中，なんです.

14. これは data の平均値を (今のところ特に理由もなく) 使っています.

```
> y <- 0:9
> prob <- dpois(y, lambda = 3.56)
> plot(y, prob, type = "b", lty = 2)
```




```
> data.frame(y, prob) # y と prob を表にしてみる
   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
```

上の図表で示そうとしているコトは何なのかを説明してみると,

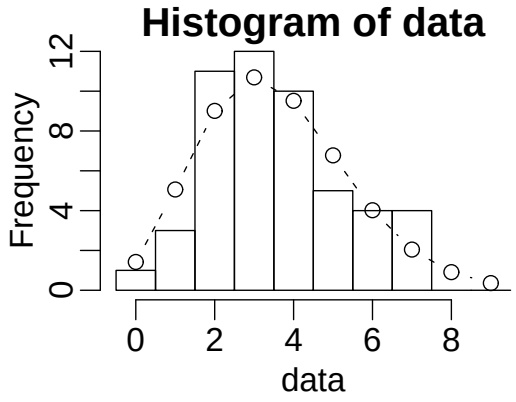
- ポアソン分布は，一日にくるメイル数 y (この y は $y \in \{0, 1, 2, 3, \cdots\}$) と，y それぞれに対応する確率 (ここでは prob) の組みあわせである
- R では「一日 y 通くる確率」である prob は dpois(y, lambda = 平

均値) で計算できる

といったことなのです. だいたいわかるでしょうか? この平均が 3.56 であるポアソン分布だと, 一日のメイル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日に 3 通ぐらいくる場合で, その確率は 0.21 ぐらい, ということをあらわしています. 平均 3.56 のポアソン分布なるものが「なんとなく」わかりましたか? [15]

ここで重要なのは, 統計モデリングにおいてはこの確率分布 (平均 3.56 のポアソン分布) を使って観測データ (data, つまり一日にくるメイル数, 50 日ぶん) にみられるパターンが「説明」できる, と考えていることです. それは図で示すとこうなります. [16]

```
> # data, y, prob はすでに定義ずみだとします
> hist(data, breaks = seq(-0.5, 8.5, 1))
> lines(y, prob * 50, type = "b", lty = 2) # 50 日ぶん
```



ヒストグラムで示されている観測データと, ポアソン分布による予測がかなり「近い」というかんじですよね? こういう関係をみて, [17]

- メイル来信数の観測データがポアソン分布 (平均 3.56 の場合) の予測とだいたい「あって」いるようにみえる
- このポアソン分布を使った統計モデルによって観測された現象が「説明」 (あるいは「表現」) されてるよね

とみなしてしまうのが統計モデリングなのです.

15. なぜ「平均 3.56」という数値を使うのか? それはこのあとの最尤推定の節で説明します.

16. ここでは hist() でヒストグラムを表示させてから, それを消さずに lines() によってポアソン分布による 50 日ぶんの予測値を上がきしています. この「上がき」技法は R による作図のたいへん重要な基本わざです.

17. もちろん実際の統計モデリングはもう少し「客観性」ごときものが要求されます.「ポアソン分布はいいのかもしれないけど, どうして平均 3.56 なの?」「他の分布はありえないわけ?」「『あっている』って何をもってそういうわけ?」……こういった疑問に対応できる方向で, この講義をススめていきたい, と考えています.

2. ポアソン分布とは何か?

ここまでで，観測データと確率分布を使った統計モデルの関係がわかってきた，ということにしましょう [18] ……上の「メイル数データ」の統計モデリングに使った確率分布，すなわちポアソン分布の性質をもう少しくわしく調べてみましょう．

ポアソン分布がどんなものであるかを規定している確率密度関数 (probablistic density function) は

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

と定義され，左辺の $p(y \mid \lambda)$ は [19] (平均値をあらわす) パラメーター λ (ラムダ) がある値のときに，確率変数が y (例: メイル数が 2 のときは $y = 2$ と考える) である確率，という意味です．その確率というのが右辺で定義されていて，y とパラメーター λ だけで定義されています． [20]

ついでに説明しとくと $\exp(-y) = e^{-y}$ で e は自然対数の底 $(e = 2.7182\cdots)$ です．$y!$ 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．

さて，ポアソン分布の性質をいくつかあげてみると……

- $y = 0, 1, 2, 3, \cdots, \infty$ (無限大) の値をとる，$\sum_{y=1}^{\infty} p(y \mid \lambda) = 1$
- 分布の平均は λ である $(\lambda \geq 0)$
- 分布のカタチを決めるパラメーターは λ だけである
- 分散と平均値は等しい: $\lambda =$ 平均 $=$ 分散

最後に「分散」なる用語がでてきました．とりあえず，これは「値のちらばりぐあい」とでも考えてください． [21]

ポアソン分布のパラメーター λ を変化させると，分布のカタチはこのように変化していきます． [22] [23]

```
> y <- 0:20
> plot(y, dpois(y, lambda = 3.5), type = "b", lty = 2, pch = 21,
+ ylab = "prob")
> lines(y, dpois(y, lambda = 7.7), type = "b", lty = 2, pch = 23)
> lines(y, dpois(y, lambda = 15.1), type = "b", lty = 2, pch = 24)
> legend("topright", legend = c(3.5, 7.7, 15.1), pch = c(21, 23, 24),
+ title = "lambda", cex = 0.7)
```

(次のページにつづく)

18. だいじょうぶかな?

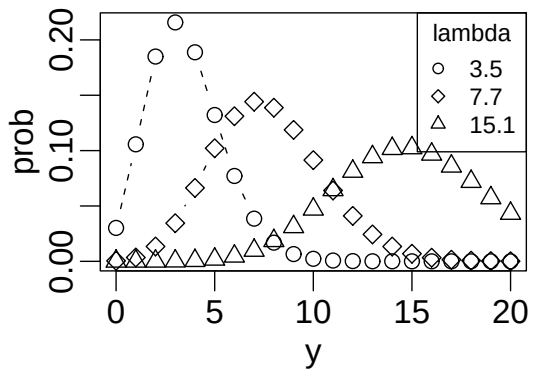
19. めんどうなときは単に $p(y)$ と書かれたりします.

20. なンでこういう式になるのか? 興味があるヒトはぜひ統計学本などで調べてみてください.

21. またあとで詳しく説明する機会があるかもしれません…… じつはこの講義では平均とは何かについてもきちんと説明してませんね!

22. ここでは plot() で「わく」を作らせて，さらに lines() による上がき，というわざが使われています

23. R のコマンドプロンプトでカッコを閉じないで改行すると，次の行頭に > ではなく + があらわれて，前の行のつづきを書くことができます.



ここで統計モデルの部品としてこのポアソン分布が選ばれた理由は

- data に含まれてる値 (y としましょうか) が $0, 1, 2, \cdots$ といった非負の整数である
- y に下限 (ゼロ) はあるみたいだけど上限はよくわからない
- 標本平均と標本分散 [24] がだいたい等しい

24. 標本平均・標本分散とは，それぞれ手もちの観測データの平均値と分散のことです.

```
> mean(data) # data の標本平均を計算
[1] 3.56
> var(data) # data の標本分散を計算
[1] 2.986122
```

といった理由によるものです.

ある確率論的な事象 (event) がポアソン分布になる，というのはどういう条件のもとでおこることなのでしょうか? たとえば，ここであつかってるような「一日にくるメイル数」だとすると，メイル数がポアソン分布になる必要条件の例としては

- 毎日くるメイル数は一定— 曜日・季節による変化があってはいけない
- 毎日くるメイル数は独立— たとえば昨日きたメイル数に依存してはいけない
- 一日の中でも各メイルの来信は独立— ある日はメイルやりとりが頻繁だから増えた，となってはいけない

といったことがあげられます. これらの条件が満たされてないときはどうしたらよいのでしょうか? 対策はふたつあります.

1. 上の条件を満たしてないみたいだけど，現象の**近似的な表現手段**としてポアソン分布を使う
2. ポアソン分布では観測された現象が説明できないから，**ポアソン分布を少し修正**した確率分布を使った統計モデリングにする

対策 2. の例としては，たとえばポアソン分布の平均値が日によって変わっているとしましょう．そうするとメイル数の標本分散は標本平均よりかなり大きくなります．こういう場合は**負の二項分布** (negative binomial distribution) を使った統計モデルや一般化線形混合モデル [25] を使ってメイル数を表現する，という方法もあります．

対策 1. と 2. のどちらをすればよいのでしょうか? これは状況によるのであり，判断するためにはそのデータについていろいろと調べなければなりません．データがどのようにして得られたのか，どのようにして生成されていると考えられるデータなのか，データの標本平均と標本分散はどういった関係にあるのか，といったことをよく考え R でデータをよく調べる必要があります．ここはデータ解析者の統計モデリングの**技術とセンス**が問われるところです． [26]

25. この講義の第 6 回であつかいます……予定でしたが第 6 回の講義はどこかに消えてしまいました．

26. そしてこのあとの講義で述べる「モデル選択」を使えば，どっちの対策が良いのかある程度は「客観的」に判断できるかもしれません．

3. 統計モデルの中核概念: 乱数発生と推定

ここまでで「ポアソン分布ってのは何となくこんなものらしい」というのがわかったコトにして，つぎにこういった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考えてみましょう．

私たちが何か観測データ，たとえば先ほどから使ってる `data`

```
> 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
```

こういう数字の羅列を見たときに,「ああ，これは何か確率論的な現象で統計モデリングによって記述可能だな」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 2 のように考えている，ということになります．研究対象である現象 (生態学であつかうのはおもに自然現象ですが) の実体は確率分布である， [27] と．私たちはこれを観測することによって，乱数 (random number) の集まりを標本 (sample) として得ている，と考えます．確率分布から乱数をとることを **sampling** とよびます．

27. まあ，人間にはなぜだかよくわからない理由で

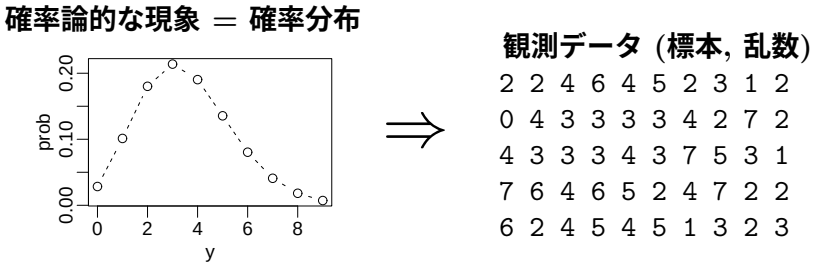


図 2: 確率分布から乱数発生，というモデル

逆に，観測データ (標本) から「そもそもこれを生み出した確率分布はどんなものなんだ?」を決めてみようとするのが**推定** (estimation) です (図 3).

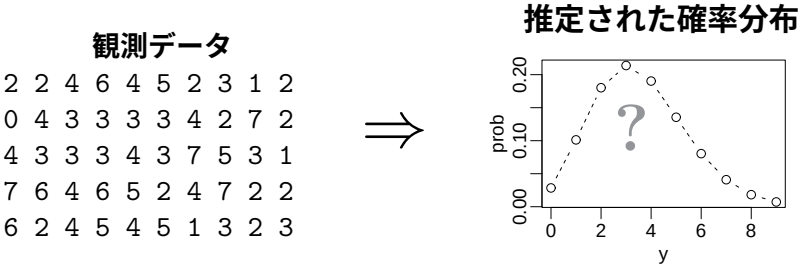


図 3: 観測データから確率分布を推定する

このように

1. ばらつきのあるデータ (標本) はある確率分布 X から生成されたのだろう (自然現象ってのは確率分布 X なのだろう)
2. 私たちは確率分布 X がどういうカタチをしているのか知らないけれど，データから確率分布 X のカタチが推定できるのではないかな

と考えることが統計モデリングの中核的な考えかた，となります．確率分布のカタチは**パラメーター** (parameter) に依存しており，[28] その値をデータから決めてやることをパラメーター推定ともいいます．このことから

- 現象がどのような確率分布で説明されそうか
- データからどのようにパラメーターの値を推定するか

この点をはっきりさせることが自然科学のデータ解析にとっては重要になります．[29] しかしながら，多くの自然科学の論文ではこの点がかなりいいかげんにあつかわれています．つまりきちんと説明されていませんし，おそらくよく考えられていないのでしょう．

この節の最後に，今回とりあげている架空データである `data` つまり「毎日くるメイル数，50 日ぶん」の生成方法について説明してみましょう．こ

28. ところで世の中では「ぱらめとりっく検定」といえば正規分布を仮定した統計モデルの検定のこと，てな意味で使われてしまっていたりしますが，これはまったく不適切なものです．どんな確率分布だってパラメーターによってカタチを決められているわけですから．

29. 皆さんが大好きな「検定」なんてのは，これらに比べると瑣末なことだと私は考えています．検定なるものは推定されたパラメーターをどう比較するかという検討なので，統計モデリングや推定をきちんとしてはじめて議論可能となるからです．

れはホントの観測データではなく，私が R の乱数生成機能をつかって作ったものです．R にはじつにさまざまな乱数生成機能がついています．たとえばポアソン分布にしたがう乱数を生成するのが rpois() 関数で，data はこれを使って生成されました．

```
> data <- rpois(50, lambda = 3.5)
> 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
```

じつは「真の平均値」は 3.56 ではなく 3.5 だったのです．このように観測されたデータから推定された平均値と「真の平均値」は異なります．

乱数生成のおもしろいところは結果が毎回毎回かわるところです．たとえば，

```
> y <- rpois(50, lambda = 3.5); print(y); print(mean(y))
 [1] 5 6 2 5 2 4 3 4 4 6 4 1 3 5 5 1 3 1 5 2 0 2 0 4 3
[26] 2 6 0 2 3 6 3 2 3 1 2 6 4 2 2 3 4 0 6 5 5 4 1 5 5
[1] 3.24
```

ここでは標本平均値もいっしょに計算していて，上の「データ」だと標本平均値は 3.24 になりました．あと二回ほど同じことをやってみましょうか．

```
> y <- rpois(50, lambda = 3.5); print(y); print(mean(y))
 [1] 1 4 2 2 4 1 3 1 2 5 4 3 5 7 1 1 2 2 3 9 0 3 5 3 5
[26] 3 3 3 4 4 4 4 3 0 4 9 3 1 7 3 2 1 2 4 3 0 6 2 2 2
[1] 3.14

> y <- rpois(50, lambda = 3.5); print(y); print(mean(y))
 [1]  5  7  2  2  4  3  5  4  4  3  3  7  5  3  5  2  5
[18]  6  2  4  4  2  3  9  1  6  0  3  3 11  1  2  2  3
[35]  3  3  2  2  3  4  3  4  6  5  4  4  4  4  4  2
[1] 3.76
```

2 回目・3 回目の標本平均値はそれぞれ 3.14 と 3.76 になっていますね．

4. ポアソン分布のパラメーター λ を最尤推定

ここまでメイル来信数の架空データである data の解析として，

1. ばらつきのあるデータ data はポアソン分布から生成されたのだろう
2. 私たちはそのポアソン分布がどういうカタチをしているのか知らないけれど，data の標本平均値 3.56 を使うと (データを生成したもともとの) ポアソン分布のカタチを決めているパラメーター λ [30] が推定できるのではないかな

というふうに統計モデリングしてきました．1. でポアソン分布を選んだ理由は上で述べてみました．それでは 2. のように「標本平均値 3.56 がなんとなくホントの平均値に『近い』のではないかな?」と考えた理由は何でしょうか? このあたりは**最尤**(さいゆう)**推定法** (maximum likelihood estimation) の考えかたにもとづいています． [31]

メイル来信数データである data の第 i 日目の来信数を y_i としましょう，つまり $\{y_1 = 2, y_2 = 2, y_3 = 4, \cdots, y_{49} = 2, y_{50} = 3\}$ ということです．このときに「平均 λ のポアソン分布から $\{y_i\} = \{y_1, y_2, y_3, \cdots, y_{50}\}$ が得られる確率」のことを尤度(ゆうど)(likelihood) とよびます．ここでポアソン分布を

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

とすると尤度 (尤度関数) は [32] [33]

$$\begin{aligned} L(\lambda \mid \{y_i\}) &= (y_1\text{が } 2 \text{ である確率}) \times (y_2\text{が } 2 \text{ である確率}) \\ &\quad \times \cdots \times (y_{50}\text{が } 3 \text{ である確率}) \\ &= p(y_1 \mid \lambda) \times p(y_2 \mid \lambda) \times p(y_3 \mid \lambda) \times \cdots \times p(y_{50} \mid \lambda) \\ &= \prod_{i=1}^{50} p(y_i \mid \lambda) \\ &= \prod_{i=1}^{50} \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} \end{aligned}$$

となります．このように定義された「データが得られる確率」こと尤度，つまり**尤**(もっとも)らしさを最大化する λ が「一番よい」λ と考えることにしましょう．

尤度 $L(\lambda \mid \{y_i\})$ を最大化する λ とはどのようなものなのかを調べてみましょう．尤度がカケ算のカタチだとたいへん扱いずらいので，対数変換した対数尤度 (log likelihood) あるいは対数尤度関数になおして計算します． [34]

$$\log L(\lambda \mid \{y_i\}) = \sum_{i=1}^{50} \left\{ y_i \log(\lambda) - \lambda - \sum_{k=1}^{y_i} \log(k) \right\}$$

ここで λ を変化させていって対数尤度 $\log L(\lambda \mid \{y_i\})$ が最大になる点があるだろう，と考えます．このように対数尤度を最大化する λ を $\hat{\lambda}$(ラムダハット) とし

30. ポアソン分布の λ は平均・分散をあらわすパラメーターでしたね

31. 最尤法という場合もあります.

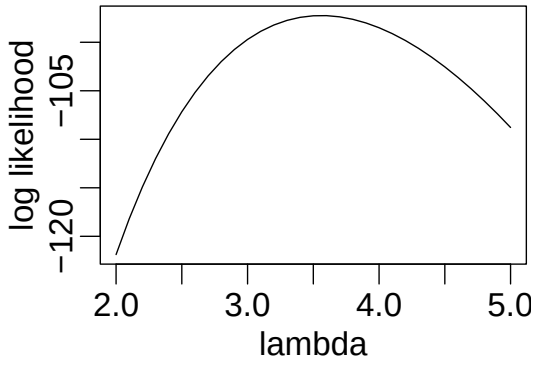
32. この式は「平均 λ のポアソン分布のもとで観測データが得られる確率」の意味ですが，形式的には $L(\lambda \mid \{y_i\})$，つまり「ある観測データのもとでパラメーター λ が得られる確率」というふうに尤度はパラメーターの関数であるように定義しています.

33. $\prod$ は積の記号です.

34. $\log(L)$ は L の単調増加関数なので，L の最大化 $\Leftrightarrow$ $\log(L)$ の最大化，となっています.

ます．ここで data が決まっているときの，λ と対数尤度の関係を図にしてみましょう．

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



この図をみると，対数尤度 (縦軸) が最大になるふきんでは対数尤度関数の「傾き」はゼロになっていることがわかります．つまり，対数尤度関数を横軸 λ で偏微分したときに

$$\frac{\partial \log L(\lambda \mid \{y_i\})}{\partial \lambda} = \sum_{i=1}^{50} \left\{ \frac{y_i}{\lambda} - 1 \right\} = \frac{\sum_{i=1}^{50} y_i}{\lambda} - 50$$

この $\log L(\lambda \mid \{y_i\})$ がゼロになるような λ が $\hat{\lambda}$ なのです．実際に計算してみると

$$\frac{\sum_{i=1}^{50} y_i}{\hat{\lambda}} - 50 = 0$$

なので

$$\hat{\lambda} = \frac{\sum_{i=1}^{50} y_i}{50} = \frac{\text{全部の } y_i \text{の和}}{\text{データ数}} = \text{データの平均値}$$

となっています．[35] このように $\{y_i\}$ に具体的な値をいれない状態の $\hat{\lambda}$ のことを **最尤推定量** (maximum likelihood estimator)，さらに $\{y_1 = 2, y_2 = 2, y_3 = 4, \cdots, y_{49} = 2, y_{50} = 3\}$ というふうに具体的に y_i に値をいれて計算して得た $\hat{\lambda} = 3.56$ のことを **最尤推定値** (maximum likelihood estimate) と言います．

つまりこのような単純な状況でポアソン分布を仮定したときに，data という観測データのもとでは，最尤推定すると平均値 3.56 が最も尤もらしい

35. 残念ながら，パラメーターまわりがもっと複雑になってしまう実際のデータ解析では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た対数尤度最大化によって最尤推定値を数値的にもとめます．次回以降で紹介する一般化線形モデルの推定関数などではこの数値的な最尤法によってパラメーターを推定しています．

値である，ということがわかりました．[36]

36. 尤度と似て非なるものがあります．準尤度 (quasi likelihood) は overdispersion を表現するために作られ，擬似尤度 (pesudo likelihood) は空間相関など単純な最尤法ではあつかいにくい現象のモデリングのために作られました．どちらも現在はあまり使われていません．Bayesian な方法が普及したので，これらが不要になったからです．

5. 二項分布で「あり・なし」データをモデル化

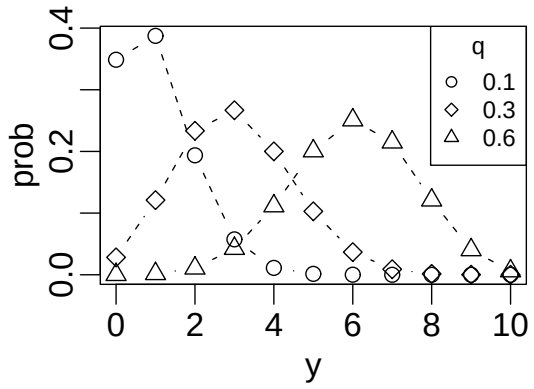
データ数が 0 個, 1 個, 2 個, $\cdots$ というふうに数えられるカウントデータの統計モデリングにつかえる確率分布はポアソン分布のほかにもいろいろとあります．

ポアソン分布はデータが 0 以上のどんな値でもとりうることをあらわしている確率分布ですが，そうではないカウントデータもあります．たとえば N 個のコインを投げたときに y 枚がおもてになった，という観測データでは y が $\{0, 1, 2, \cdots, N\}$ の値しかとることができません．このときには **二項分布** (binomial distribution)

$$p(y \mid N, q) = \binom{N}{y} q^y (1-q)^{N-y}$$

を使うと y が $\{0, 1, 2, \cdots, N\}$ になる確率を計算できます．

```
> y <- 0:10
> plot(y, dbinom(y, 10, prob = 0.1), type = "b", lty = 2, pch = 21,
+ ylab = "prob")
> lines(y, dbinom(y, 10, prob = 0.3), type = "b", lty = 2, pch = 23)
> lines(y, dbinom(y, 10, prob = 0.6), type = "b", lty = 2, pch = 24)
> legend("topright", legend = c(0.1, 0.3, 0.6), pch = c(21, 23, 24),
+ title = "q", cex = 0.7)
```



この二項分布を使った統計モデルは，たとえば生態学のデータ解析では，たとえばある実験処理をして N 個体中 y 個体が応答したといった，反応の「ある・なし」データ解析に使います．

二項分布については次の次の講義「第 4 回: 11/10 (月) 一般化線形モデル (GLM) 2」でくわしく紹介します．

6. 正規分布についてもちょっとばかり

よく理解されないままよく使われてしまっている**正規分布** (normal distribution; 別名: ガウス分布, Gaussian distribution) [37] はポアソン分布や二項分布とは異なり，$\{1.5, -3.2, 7.7, \cdots\}$ といった**連続値** のデータをあつかう確率分布です．正規分布で 0 個，1 個，2 個，$\cdots$ といったカウントデータをあつかっているヒトがいますけれど，これはまったく不適切な使いかたです．

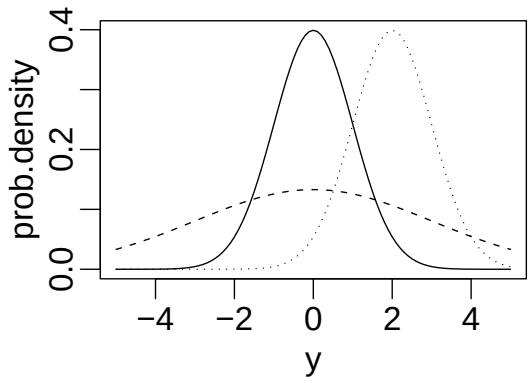
37. というか,「なんでもかんでも正規分布」なヒトたちはほとんど「確率分布とは何か」といったことを考えたことがないのでは? と思えてしまいます.

正規分布はポアソン分布と同じく平均値のパラメーターをもちます．平均をあらわすパラメーター $\overset{ミュー}{\mu}$ は $\pm\infty$ の範囲で自由に平均値を変えることができます． [38] ．またポアソン分布とは異なり，値のばらつきをあらわす標準偏差 [39] をあらわすパラメーター $\overset{シグマ}{\sigma}$ はゼロ以上なら自由な値を設定できます．つまり，確率分布の平均だけでなく，ばらつき具合も指定できます．

38. ポアソン分布では平均値パラメーター λ は非負でなければならない, という制約があります.

39. 標準偏差 ^{2} = 分散

```
> y <- seq(-5, 5, 0.1)
> plot(y, dnorm(y, mean = 0, sd = 1), type = "l",
+ ylab = "prob.density") # 実線
> lines(y, dnorm(y, mean = 0, sd = 3), lty = 2) # 破線
> lines(y, dnorm(y, mean = 3, sd = 1), lty = 3) # 点線
```



正規分布みたいな連続値の確率分布にはちょっと面倒なところがあって，上のグラフは確率ではなく確率密度をあらわしています．つまり正規分布の確率密度関数は平均パラメーター μ と標準偏差パラメーター σ を使って

$$p(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-\mu)^2}{2\sigma^2})$$

というふうに定義されるのですが，これは確率ではなく確率密度なのです．

どういうことかというと……たとえば成人男子の平均身長が 170 cm でその標準偏差が 5 cm だとします． [40] このときに「ある人の身長が 177 cm から 178 cm にある確率を計算せよ」という問題には

40. データにもとづく値ではありません.

```
> pnorm(178, mean = 170, sd = 5) - pnorm(177, mean = 170, sd = 5)
[1] 0.02595737

> dnorm(177.5, mean = 170, sd = 5) * 1.0 # 近似的な確率の簡単計算法
[1] 0.02590352
```

と答えることができます．しかし「ある人の身長が 178 cm ぴったりである確率を計算せよ」という問題だと

```
> pnorm(178, mean = 170, sd = 5) - pnorm(178, mean = 170, sd = 5)
[1] 0
```

確率はゼロになります．ただし確率密度はゼロではありません．

```
> dnorm(178, mean = 170, sd = 5)
[1] 0.02218417
```

正規分布の最尤推定について少しだけ言及してみます．前回の授業で「最小二乗法は最尤推定の一部である」とのべました．最小二乗法は統計モデルの部品として正規分布を使っているときに使えるパラメーター推定法です．この関係について少し調べてみましょう．

たとえばある人間の集団の身長データを $\{y_i\}$ $(i = 1, 2, \cdots, N)$ とすると，正規分布をつかった統計モデルの尤度関数は

$$\begin{aligned} L(y_i \mid \mu, \sigma) &= \prod_{i=1}^{N} p(y \mid \mu, \sigma)\Delta y \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \mu)^2}{2\sigma^2})\Delta y \end{aligned}$$

となります．ここで Δy は「身長に関する一定の微小な幅 (例: 0.1 cm など)」 [41] です．というのも正規分布の確率密度 $p(y \mid \mu, \sigma)$ はそれだけでは「確率」ではなく「(身長の) 幅」をかけて「面積」にしてやらないと確率になりません． [42]

対数尤度関数は

$$\log L(y_i \mid \mu, \sigma) = -0.5N \log(2\pi\sigma^2) - \frac{\sum_{i=1}^{N}(y_i - \mu)^2}{2\sigma^2} + N\Delta y$$

いま標準偏差パラメーターである σ が μ とは無関係ななにか定数だとすると， [43] 対数尤度 $\log L(y_i \mid \mu, \sigma)$ を最大化するにはうしろの項つまり $-\frac{\sum_{i=1}^{N}(y_i - \mu)^2}{2\sigma^2}$ の分子 $\sum_{i=1}^{N}(y_i - \mu)^2$ を最小に (「二乗誤差の和」を最小

41. 小さければなんでもいいです．

42. 先ほどの drnorm() などを使った正規分布の確率計算を参照してください．

43. もちろん $N\Delta y$ も定数です．

化) するような平均のパラメーター $\hat{\mu}$ を推定してやればよい，ということがわかると思います.

このことから正規分布を部品とする統計モデルにおける最尤推定法は最小二乗法，[44] が成立しています.

44. ただし σ が一定，といった仮定が必要ですね.

7. 統計モデルにどの確率分布を使えそうか

観測データを説明する統計モデル，その基本部品である確率分布はどう選べばよいのでしょうか? 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.

1. 説明したい量は**離散**か**連続**か?
   - 離散: { 生きてる, 死んでる }，カウントデータ, $\cdots$
   - 連続: $\{0.56, 1.33, 12.4, 9.84, \cdots\}$，$\cdots$
2. 説明した量の**範囲**は?
   - $\{0, 1, \cdots, N\}$, $\{0, 1, \cdots, \infty\}$, $[y_{\min}, y_{\max}]$, $[-\infty, \infty]$, $\cdots$
3. 説明したい量の標本分散 (**ばらつき**) と標本平均の関係は?
   - 分散 $\approx$ 定数, 分散 $\approx$ 平均, 分散 $\propto$ 平均, 分散 $\propto$ 平均^n, $\cdots$

いくつかの確率分布について，それが統計モデルの部品として使えそうな必要条件をざっとみてみると，

- **ポアソン分布**: データが離散値，ゼロ以上の範囲，上限とくになし，標本平均 $\approx$ 標本分散
- **負の二項分布**: データが離散値，ゼロ以上の範囲，上限とくになし，標本平均 $<$ 標本分散
- **二項分布**: データが離散値，ゼロ以上で有限の範囲 ($\{0, 1, 2, \cdots, N\}$)
- **正規分布**: データが連続値，範囲が $[-\infty, +\infty]$
- **対数正規分布**: データが連続値，範囲が $[0, +\infty]$
- **ガンマ分布**: データが連続値，範囲が $[0, +\infty]$
- **ベータ分布**: データが連続値，範囲が $[0, 1]$

また R にはさまざまな確率分布が使えるように準備されています．いくつかの関数について，それぞれの乱数生成関数とデータからパラメーターを推定する関数を示してみましょう．[45]

	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	`rbinom()`	`glm(family = binomial)`
	二項分布	`rbinom()`	`glm(family = binomial)`
	ポアソン分布	`rpois()`	`glm(family = poisson)`
	負の二項分布	`rnbinom()`	`glm.nb()`
(連続)	ガンマ分布	`rgamma()`	`glm(family = gamma)`
	正規分布	`rnorm()`	`glm(family = gaussian)`

一般化線形モデルの推定関数である `glm()` の使いかたは次回以降の講義であつかうことにします．

次回は「第 3 回: 一般化線形モデル (GLM) 1」，ポアソン分布を仮定した統計モデルを**一般化線形モデル**としてあつかう方法について考えてみましょう．

45. ベルヌーイ分布と二項分布はほとんど同じものです．二項分布は N 個の独立した事象がどれも確率 p で生じるときに，いくつの事象が生じるかをあらわす確率分布で，確率変数は $\{0, 1, 2, \cdots, N\}$ の値をとります．ベルヌーイ分布は二項分布の $N = 1$ の場合で，確率変数は $\{0, 1\}$ の値をとります．

確率分布と最尤推定

- 確率論的な現象は **確率分布**でうまく表現できそうだ
- 確率分布のカタチは **パラメーター**できまる
- パラメーターの値は，観測データにもとづいて **最尤推定法**で得られる

この講義で使う確率分布: ポアソン分布, 二項分布, 正規分布

2008–10–30 5/ 7

図 4: 本日の要点