統計モデリング入門 2014 (b)

確率分布と最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp

茨城大集中講義 http://goo.gl/2QNwgl

2014–10–01

ファイル更新時刻: 2014–09–29 16:15

今日のハナシ I

1. **例題: 種子数の統計モデリング**
   まあ，かなり単純な例から始めましょう
2. **データと確率分布の対応**
   確率分布は統計モデルの重要な部品
3. **ポアソン分布って何?**
   平均を変えると分布のカタチが変わる
4. **ポアソン分布のパラメーターの 最尤推定** (さいゆうすいてい)
   もっとももっともらしい推定?
5. **統計モデルの要点**
   乱数発生・推定・予測

本題にはいる前に
統計モデリング授業前半の
主題を
再確認しておきましょう

この授業であつかう統計モデルたち

線形モデルの発展

推定計算方法

MCMC

階層ベイズモデル

もっと自由な統計モデリングを！

一般化線形混合モデル

最尤推定法

個体差・場所差といった変量効果をあつかいたい

一般化線形モデル

最小二乗法

正規分布以外の確率分布をあつかいたい

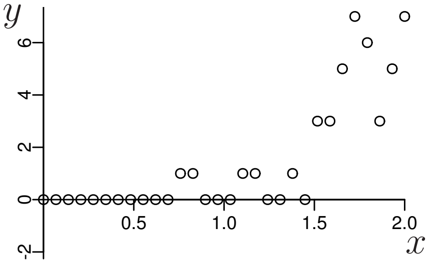
線形モデル

データの特徴にあわせて線形モデルを改良・発展させる

0 個, 1 個, 2 個と数えられるデータ

カウントデータ ($y \in \{0, 1, 2, 3, \cdots\}$ **なデータ**)

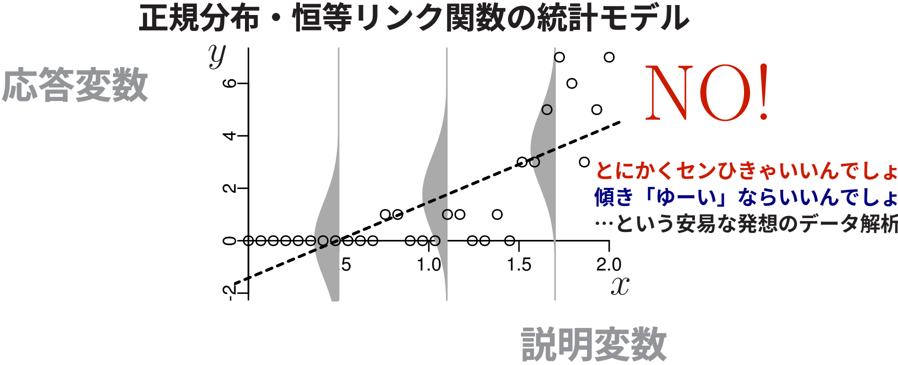
応答変数



説明変数

- たとえば x は植物個体の大きさ，y はその個体の花数
- 体サイズが大きくなると花数が増えるように見えるが……
- この現象を表現する統計モデルは?

正規分布を使った統計モデル …… ムリがある?

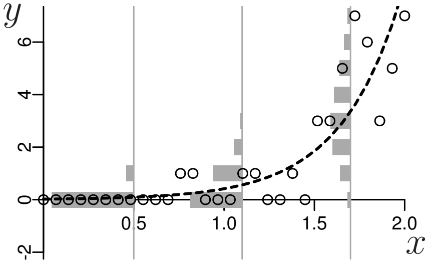


- **タテ軸のばらつきは「正規分布」なのか?**
- y **の値は** 0 **以上なのに ……**
- **平均値がマイナス?**

ポアソン分布を使った統計モデルなら良さそう?!

ポアソン分布・対数リンク関数の統計モデル

応答変数



YES!

説明変数

- タテ軸に対応する「ばらつき」
- 負の値にならない「平均値」
- 正規分布を使ってるモデルよりましだね

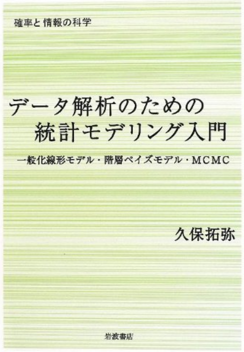
データの性質をよくみる
確率分布という部品を選ぶ
ぶらっくぼっくすにしない!

今回の内容と統計モデリング入門との対応

http://goo.gl/Ufq2

今回はおもに「第 2 章 確率分布と統計モデルの最尤推定」の内容を説明します．

- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012–05–18 刊行

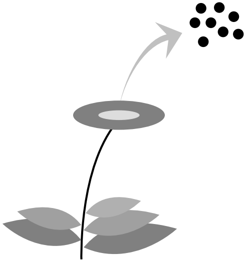


2. 例題：種子数の統計モデリング

まあ，かなり単純な例から始めましょう

R でデータをあつかいつつ

この授業では架空植物の架空データをあつかう



理由: よけいなことは考えなくてすむので

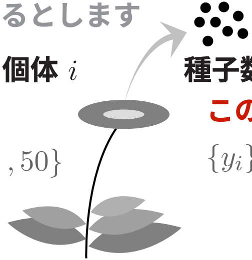
現実のデータはどれも授業で使うには難しすぎる……

こんなデータ (架空) があったとしましょう

まあ，なんだかこういうヘンな
植物を調査しているとします

個体 i

種子数 y_i



全 50 個体

$i \in \{1, 2, 3, \cdots, 50\}$

この $\{y_i\}$ が観測データ!

$\{y_i\} = \{y_1, y_2, \cdots, y_{50}\}$

このデータ $\{y_i\}$ がすでに R という統計ソフトウェアに格納されていた，としましょう

```
> 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
```

R でデータの様子をながめる

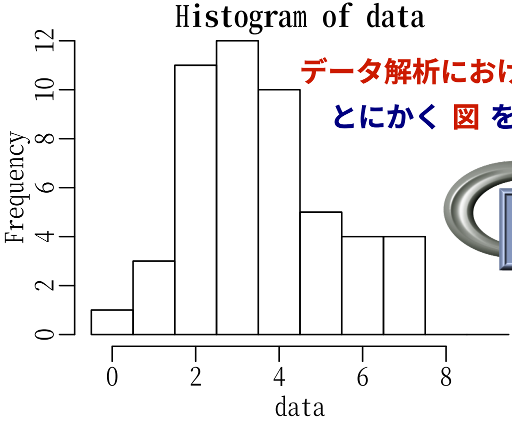
の `table()` 関数を使って種子数の頻度を調べる

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

(**種子数** 5 **は** 5 **個体，種子数** 6 **は** 4 **個体** ……)

とりあえずヒストグラムを描いてみる

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

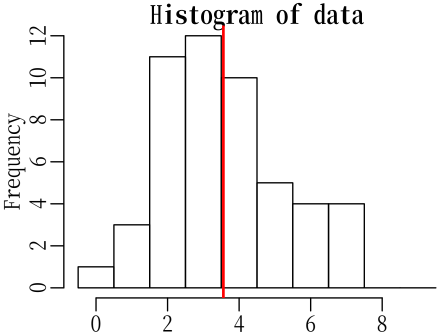


データ解析における最重要事項

とにかく 図 を描く!

標本平均という統計量

```
> mean(data)
[1] 3.56
> abline(v = mean(data), col = "red")
```



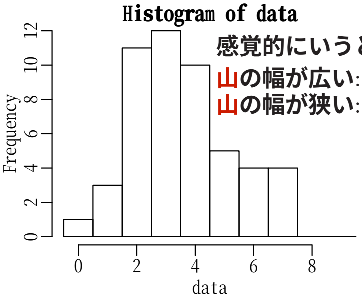
ばらつきの統計量

あるデータの **ばらつき** をあらわす標本統計量の例: **標本分散**

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

標本標準偏差 とは標本分散の平方根 $(\mathrm{SD} = \sqrt{\mathrm{variance}})$

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



感覚的にいうと
山の幅が広い: 分散が大きい
山の幅が狭い: 分散が小さい

3. データと確率分布の対応

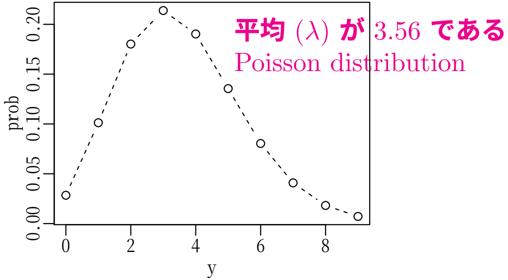
確率分布は統計モデルの重要な部品

ばらついてる**データ**を近似する道具

ポアソン分布とは何か?

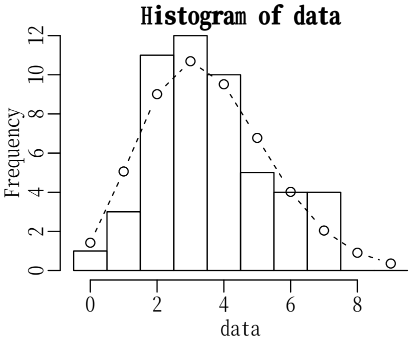
とりあえず R で作図してみる

```
> y <- 0:9                              # これは種子数（確率変数）
> prob <- dpois(y, lambda = 3.56)       # ポアソン分布の確率の計算
> plot(y, prob, type = "b", lty = 2)
```




```
> # cbind で「表」作り
> cbind(y, prob)
      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
```

データとポアソン分布を重ね合わせる




```
> hist(data, seq(-0.5, 8.5, 0.5))      # まずヒストグラムを描き
> lines(y, prob, type = "b", lty = 2) # その「上」に折れ線を描く
```

4. ポアソン分布って何?

平均を変えると分布のカタチが変わる

確率分布のカタチをきめるパラメーター

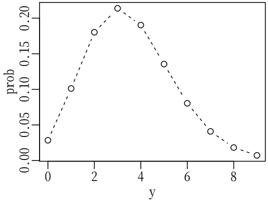
ポアソン分布を数式で表現してみる

種子数が y である確率は以下のように決まる，と考えている

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

- $y!$ **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．**
- $\exp(-\lambda) = e^{-\lambda}$ **のこと** $(e = 2.718 \cdots)$
- **ここではなぜポアソン分布の確率計算が上のようになるのかは説明しません— まあ，こういうもんだと考えて先に進みましょう**

パラメーター λ はポアソン分布の平均




```
> # cbind で「表」作り
> cbind(y, prob)
      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
```

- **平均 λ はポアソン分布の唯一のパラメーター**
- **確率分布の平均は λ である（$\lambda \geq 0$）**
- **分散と平均は等しい: $\lambda =$ 平均 $=$ 分散**
- $y \in \{0, 1, 2, \cdots, \infty\}$ **の値をとり，すべての y について和をとると 1 になる**

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

どういう場合にポアソン分布を使う?

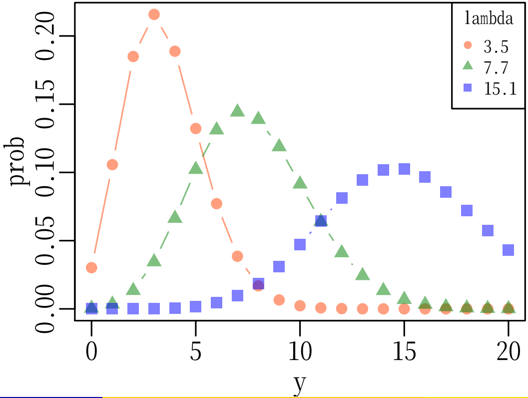
統計モデルの部品としてポアソン分布が選んだ理由:

- 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）
- 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
- この観測データでは平均と分散がだいたい等しい
  - この**だいたい等しい**があやしいのだけど，まあ気にしないことにしましょう

ポアソン分布の λ を変えてみる

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

λ は平均をあらわすパラメーター



各個体の y_i が独立にポアソン分布にしたがう

……ってどういう意味?

- 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 $y_1 = 2$ 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 $y_2 = 2$ 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 $y_3 = 4$ だった
- —（以下，同様）—

といった意味 (他個体とは無関係，ということ)

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

— (つづく)

5. ポアソン分布のパラメーターの 最尤推定（さいゆうすいてい）

もっとももっともらしい推定?

あてはめることは推定すること

尤度(ゆうど) (likelihood) とは何か?

- 最尤推定法では，**尤度**(ゆうど) という**あてはまりの良さ**をあらわす統計量に着目
- 尤度は**データが得られる確率**をかけあわせたもの
- この例題の場合，パラメーター λ を変えると尤度が変わる
- もっとも「あてはまり」が良くなる λ を見つけたい
- 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，尤度はだいたい $0.180 \times 0.180 \times 0.19 = 0.006156$ といった値になる

尤度 $L(\lambda)$ はパラメーター λ の関数

この例題の尤度:

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

尤度はしんどいので対数尤度を使う

尤度は確率 (あるいは確率密度) **の積であり，あつかいがふべん (大量のかけ算!)**

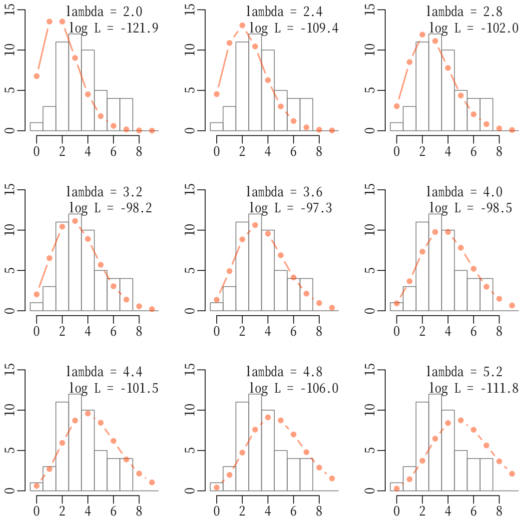
そこで，パラメーターの最尤推定では，対数尤度関数 (log likelihood function) **を使う**

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

対数尤度 $\log L(\lambda)$ **の最大化は尤度** $L(\lambda)$ **の最大化になるから**

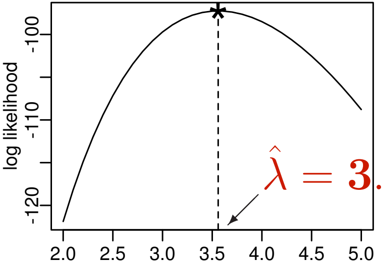
まずは，平均をあらわすパラメーター λ **を変化させていったときに，ポアソン分布のカタチと対数尤度がどのように変化するのかを調べてみましょう**

λ を変えるとあてはまりの良さが変わる



対数尤度を最大化する $\hat{\lambda}$ をさがす

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



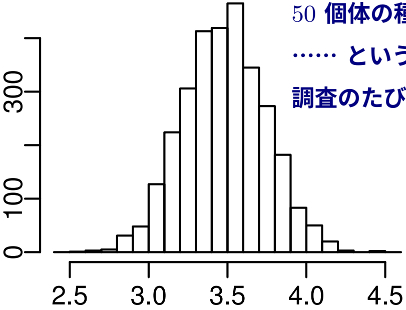
$\frac{d \log L}{d\lambda} = 0$ より

$\hat{\lambda} = 3.56$

- **最尤推定量** (ML estimator): $\sum_i y_i / 50$ 標本平均値!
- **最尤推定値** (ML estimate): $\hat{\lambda} = 3.56$ ぐらい

最尤推定を使っても真の λ は見つからない

真の λ が 3.5 の場合



試行ごとに推定された $\hat{\lambda}$

50 個体の種子数を調べる

…… ということを 3000 回くりかえし

調査のたびに $\hat{\lambda}$ を最尤推定した

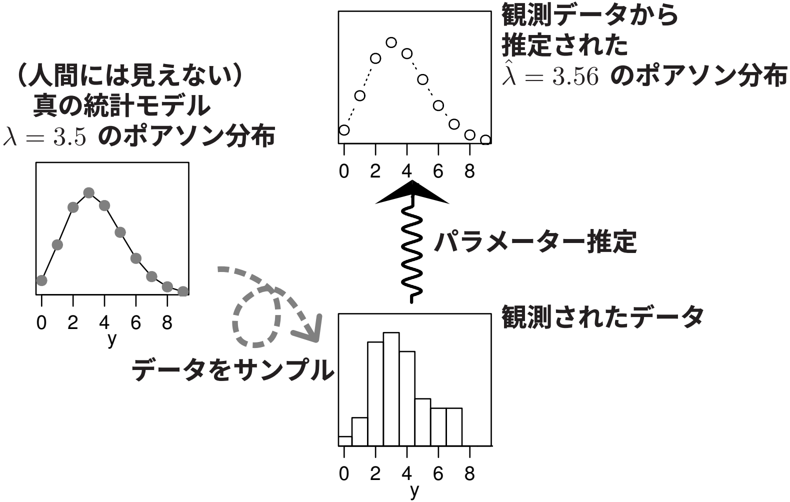
データは有限なので真の λ はわからない

6. 統計モデルの要点

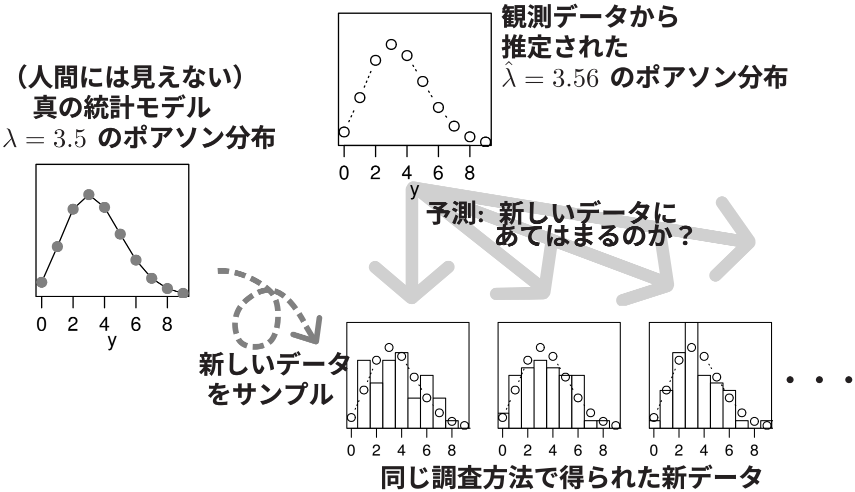
乱数発生・推定・予測

統計モデルとデータの対応づけ

統計学における推定


（人間には見えない）
真の統計モデル
$\lambda = 3.5$ のポアソン分布
0 2 4 6 8
y
データをサンプル
観測データから
推定された
$\hat{\lambda} = 3.56$ のポアソン分布
0 2 4 6 8
パラメーター推定
観測されたデータ
0 2 4 6 8
y

統計学における予測

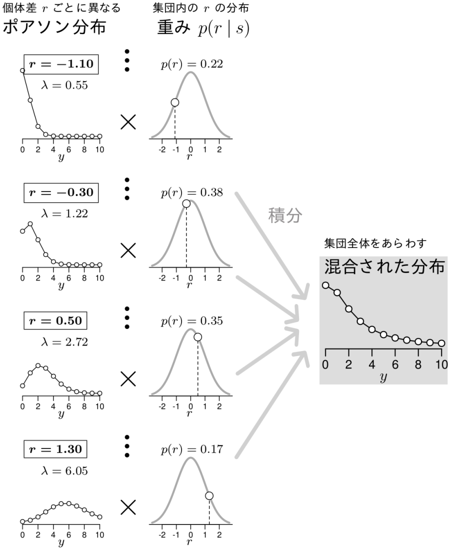

(人間には見えない)
真の統計モデル
$\lambda = 3.5$ のポアソン分布
観測データから
推定された
$\hat{\lambda} = 3.56$ のポアソン分布
予測: 新しいデータに
あてはまるのか？
新しいデータ
をサンプル
同じ調査方法で得られた新データ
0 2 4 6 8
y

この授業で登場する確率分布

- **ポアソン分布**: $y \in \{0, 1, 2, 3, \cdots\}$ **となるデータ，「y 回なにかがおこった」**
- **二項分布**: $y \in \{0, 1, 2, \cdots, N\}$ **となるデータ，「N 個のうち y 個で何かがおこった」**
- **正規分布**: $-\infty < y < \infty$ **の連続値をとるデータ**
- 一様分布，ガンマ分布 — ちょっと登場するだけ

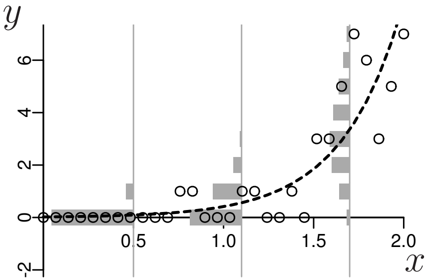
いろいろな確率分布があるけれど……

- この授業では多種多様な確率分布をあつかいません
- この授業後半では「ポアソン分布と正規分布を混ぜる」— 確率分布まぜワザを使って，現実にみられる複雑な分布を再現してみます



次回予告

The next topic



YES!

一般化線形モデルのひとつ: ポアソン回帰

Poisson Regression, a Generalized Linear Model (GLM)