『データ解析のための統計モデリング入門:
一般化線形モデル・階層ベイズモデル・MCMC』

～読書会～

# 2章 確率分布と統計モデルの最尤推定

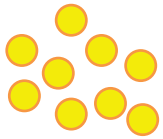
# 2 確率分布と統計モデルの最尤推定

# 確率分布　probability distribution

・統計モデルの本質的な部品
・データにみられるさまざまな「ばらつき」

例題データと確率分布の対応づけ

## 2. 1 例題: 種子数の統計モデリング



個体 $i$

種子数 $y_i$

- 0, 1, 2, … ：カウントデータ
- 非負の整数

➡ 確率分布を選ぶときの手がかり

## 2. 1 例題: 種子数の統計モデリング

Courier New

```
load("data.RData")
# ダウンロードしたデータファイルを読み込む
```

```
data
# 「data」という名前のデータの内容を表示
```

```
length(data)
# dataにはいくつのデータが含まれるのか
```

## 2. 1 例題: 種子数の統計モデリング

```
summary(data)
# dataを要約せよ
```

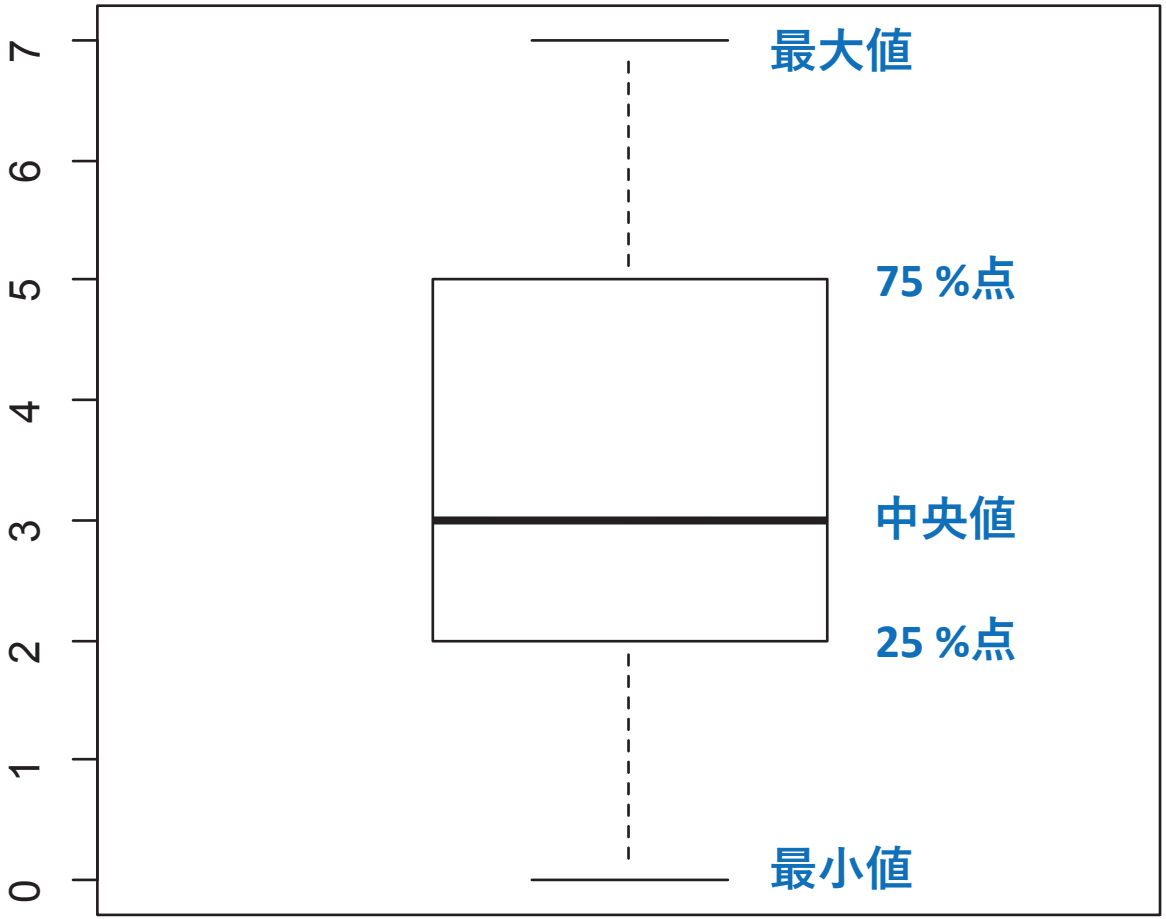
- Min.　　最小値
- Max.　　最大値
- 1st Qu.　25 %点の値
- Median　50 %点の値（標本中央値）
- 3rd Qu.　75 %点の値
- Mean　　標本平均

} 四分位数 → 箱ひげ図

**重要なのは、まず様々な方法で図示すること**

## 2. 1 例題: 種子数の統計モデリング

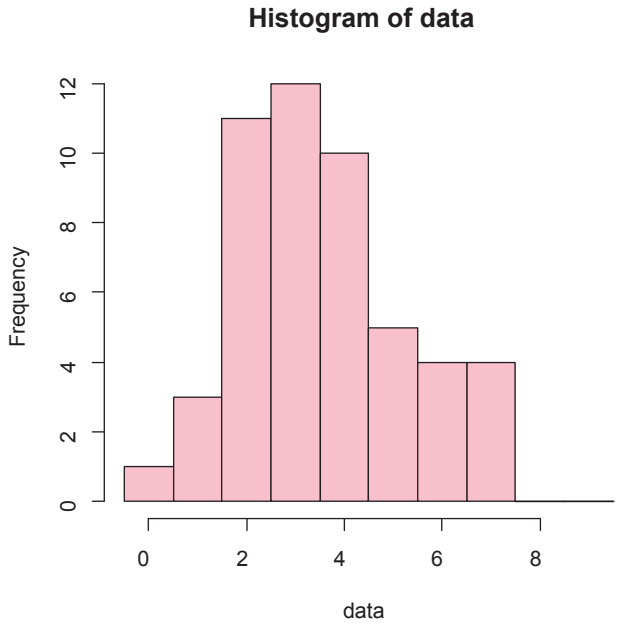
```
boxplot(data)
```



## 2. 1 例題: 種子数の統計モデリング

```
table(data)
# 度数分布表を得る
```

```
hist(data, breaks = seq(-0.5, 9.5, 1))
# ヒストグラムを図示
```



```
var(data)
# 標本分散を得る
```

標本分散 sample variance
＝データのばらつき variability

```
sd(data)
sqrt(var(data))
# 標本標準偏差を得る
```

標本標準偏差 sample standard deviation

## 2. 2 データと確率分布の対応関係をながめる

- 0, 1, 2, … ：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あり、ヒストグラムはひと山分布

➡ **ポアソン分布**（確率分布）が便利

**Poisson distribution**

# 確率分布

確率変数の値とそれが出現する確率を対応させたもの

# 確率変数 random variable

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

# 「平均3.56のポアソン分布」とは？

```
y <- 0:9
prob <- dpois(y, lambda = 3.56)
# 「平均3.56のポアソン分布にしたがって、ある個体の種子数がy個である確率」
```

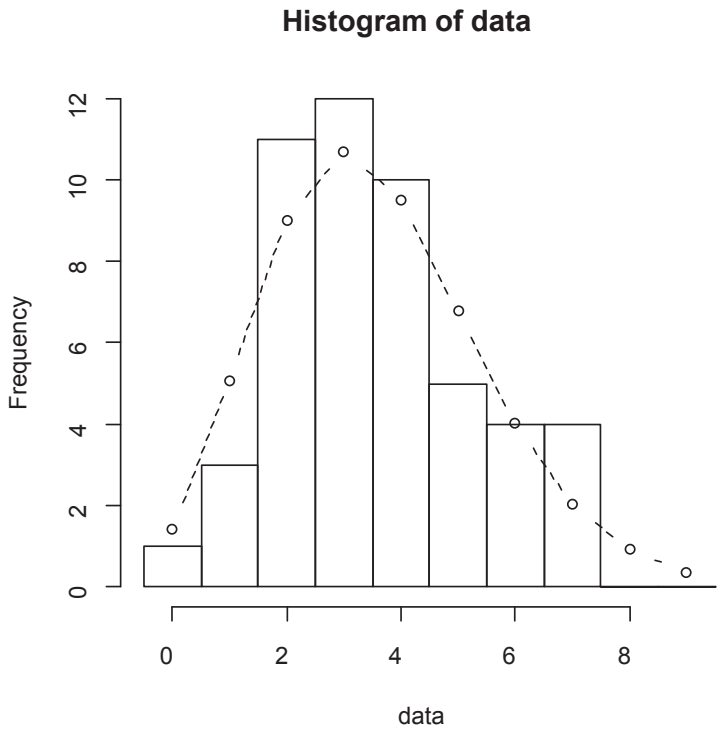
```
cbind(y, prob)
# 表形式で出力
plot(y, prob, type = "b", lty = 2)
# 図形式で出力
```

丸と折れ線

破線

## 2. 2 データと確率分布の対応関係をながめる

```
hist(data, breaks = seq(-0.5, 9.5, 1))
# ヒストグラムを図示
lines(y, 50 * prob, type =“b”, lty =2)
# 予測値（prob）を描画
```




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

# ポアソン分布の定義

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

平均が$\lambda$であるとき、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

平均$\lambda$：ポアソン分布の唯一のパラメーター

# ポアソン分布の性質

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

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

- 確率分布の平均は$\lambda$である$(\lambda \geq 0)$
- 分散と平均は等しい：$\lambda$=平均＝分散

# ばらつき＝誤差 error

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も、なんらかの理由で個体ごとに異なる種子数になっている

調査している人の間違い（測定誤差）ではない

# 最尤推定 maximum likelihood estimation

尤度（あてはまりの良さ）を最大にするようなパラメーターの値を探す

# 尤度

ある$\lambda$の値を決めたときに、すべての個体$i$についての$p(y_i|\lambda)$の積

$$L(\lambda) = \prod_i p(y_i|\lambda) = \frac{\lambda^{y_i}\exp(-\lambda)}{y_i!}$$

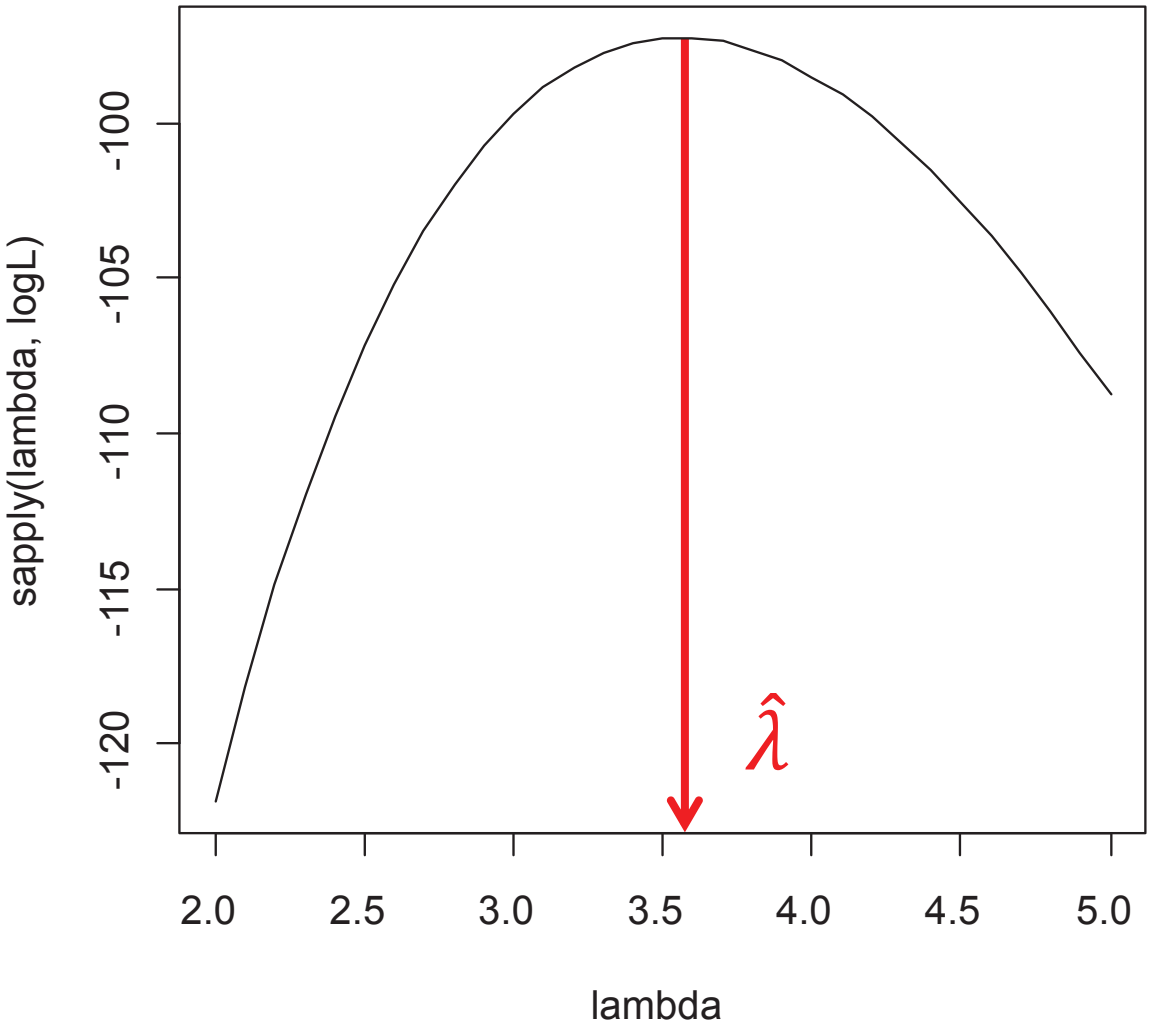
$$\log L(\lambda) = \sum_i \left( y_i \log \lambda - \lambda - \sum_k^{y_i} \log k \right)$$

# 対数尤度関数を使ってパラメータを最尤推定する

```
logL <- function(m) sum(dpois(data, m,
log = TRUE))

lambda <- seq(2, 5, 0.1)
                    イチじゃなくエル
plot(lambda, sapply(lambda, logL),
type = "l")
# 対数尤度とλの関係を図示
```

## 2. 4 ポアソン分布のパラメーターの最尤推定




**対数尤度関数が最大値で関数の傾きがゼロになる$\lambda$を探しだせばよい**

## 2. 4 ポアソン分布のパラメーターの最尤推定

### 対数尤度関数

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

### パラメーター$\lambda$で偏微分すると、

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

### これがゼロである場合、

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

## 最尤推定量 maximum likelihood estimator

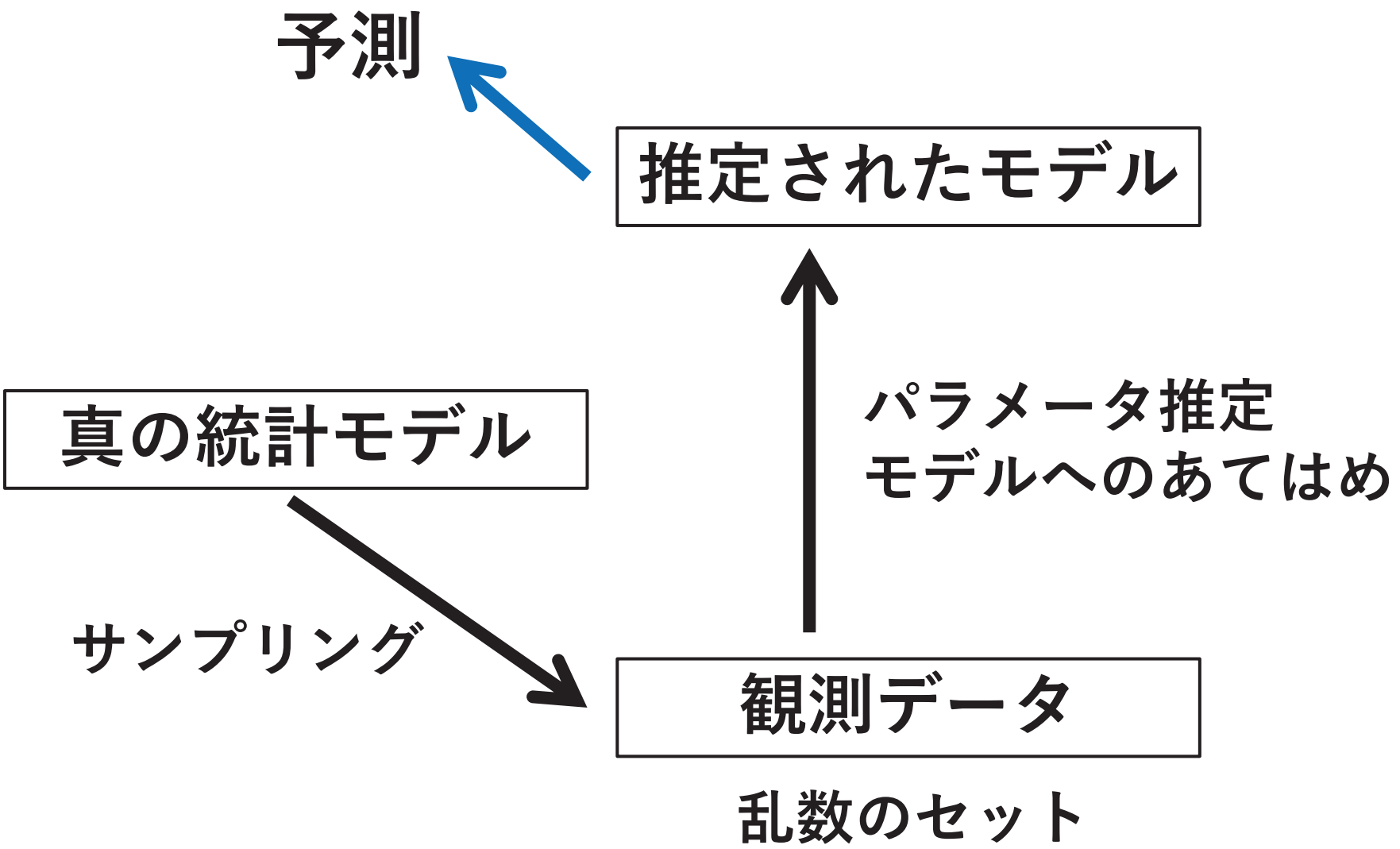
対数尤度あるいは尤度が最大になる$\hat{\lambda}$

## 最尤推定値 maximum likelihood estimate

具体的な$y_i$の値を使って評価された

$\hat{\lambda} = 3.56$

## 2. 5 統計モデルの要点：乱数発生・推定・予測



# 統計モデルを使った予測

- 次に得られる応答変数の平均だけを示す
- 次に得られるデータの予測区間も示す

時系列構造のあるデータ ➡ 将来予測

空間構造のあるデータ ➡ 欠測データの補間

# 予測の良さ goodness of prediction

推定されたモデルが新しく得られたデータにどれくらい良くあてはまるか

# 「この現象がどのような確率分布で説明されそうか」

- 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？

# 確率分布の種類

- ポアソン分布 poisson distribution
  データが離散値、ゼロ以上で上限なし、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、分散は平均の関数
- 正規分布 normal distribution
  データが連続値、範囲が[-∞, +∞]、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, +∞]、分散は平均の関数
- 一様分布 uniform distribution
  データが連続かつ有限

「ブラックボックスな統計解析」から脱するために、まず、「この統計モデリングでは、このような理由でこの確率分布を使いました」と他人にきちんと説明できるようになってください。