データ解析のための
統計モデリング入門

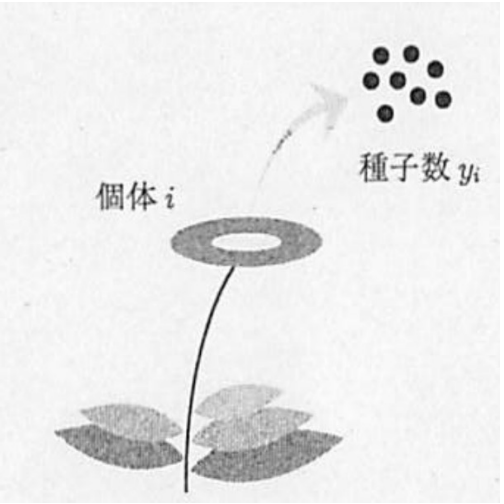
2. 確率分布と統計モデルの最尤推定

河野和平

確率分布

- データにみられるさまざまな”ばらつき”を表現
- 確率変数の値とそれが出現する確率を対応

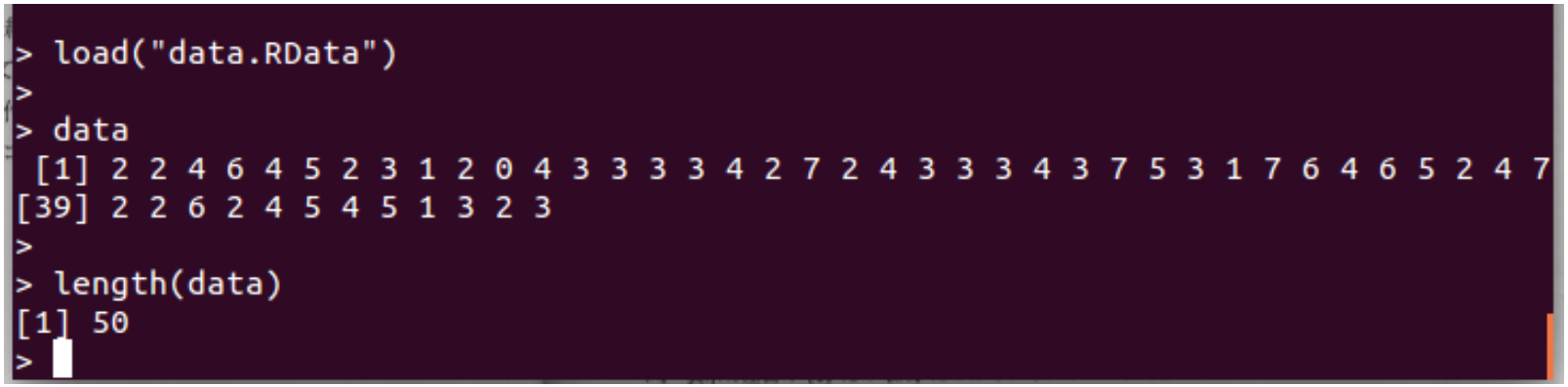
- 例題：種子数の統計モデリング



・ 個体$i \in \{1,2,...,50\}$
・ 種子数 y_i

例題データ

- 読み込み
- 表示
- 大きさ



データ解析

- 最小値, 最大値
- 昇順の25%, 50%(中央値), 75%の値
- 標本平均

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

- 分散･標準偏差

```
>
> var(data)
[1] 2.986122
> sd(data)
[1] 1.72804
>
```

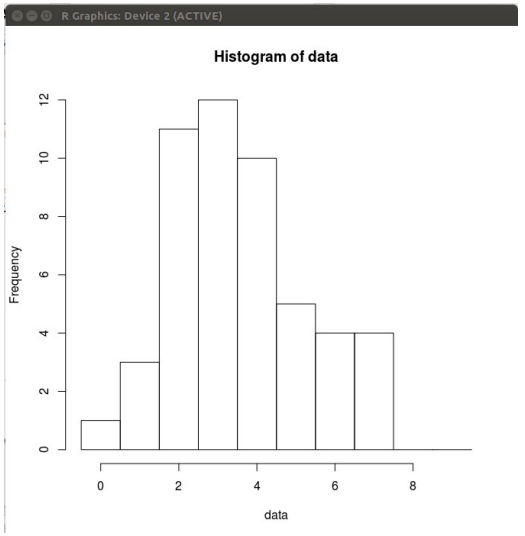
度数分布

- 度数分布

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

- ヒストグラム

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



ポアソン分布(1)

- 単位時間あたり平均 λ回起こるようなランダムな事象
- 単位時間に y回起こる確率

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

- すべての yについて和を取ると1

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

- 平均=分散= λ

ポアソン分布(2)

- 単位時間あたり λ 回起こるようなランダムな事象
- 単位時間に y 回起こる確率

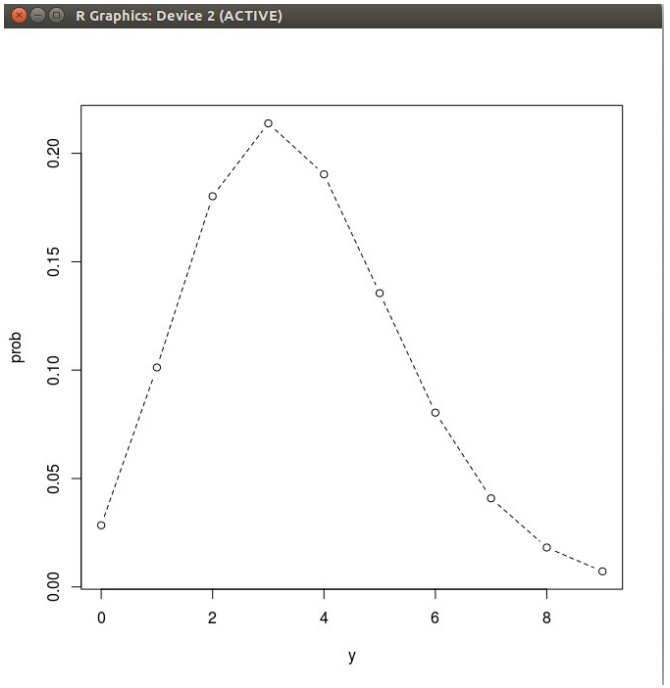
$$p(y|\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^{y}}{y!}$$

- $\lambda = 3.56$

```
> y <- 0:9
>
> y
 [1] 0 1 2 3 4 5 6 7 8 9
>
> prob <- dpois(y,lambda = 3.56)
>
> prob
 [1] 0.02843882 0.10124222 0.18021114 0.21385056 0.19032700 0.13551282
 [7] 0.08040427 0.04089132 0.01819664 0.00719778
>
```

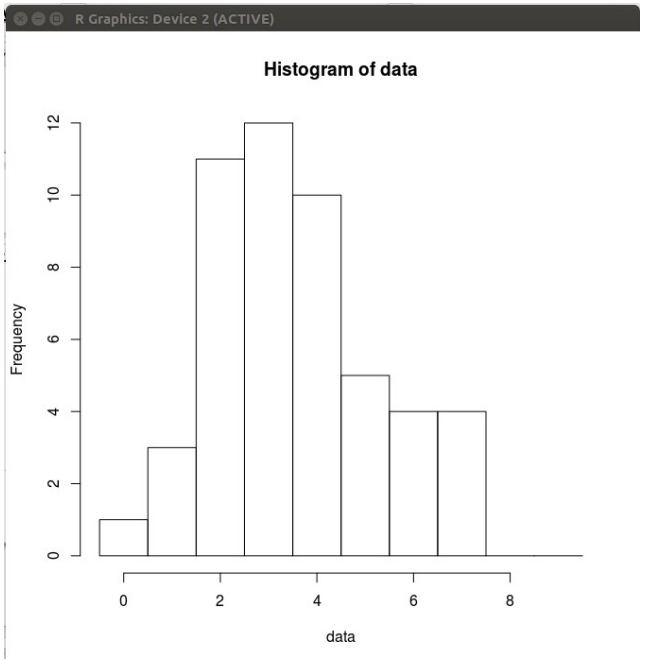
グラフ

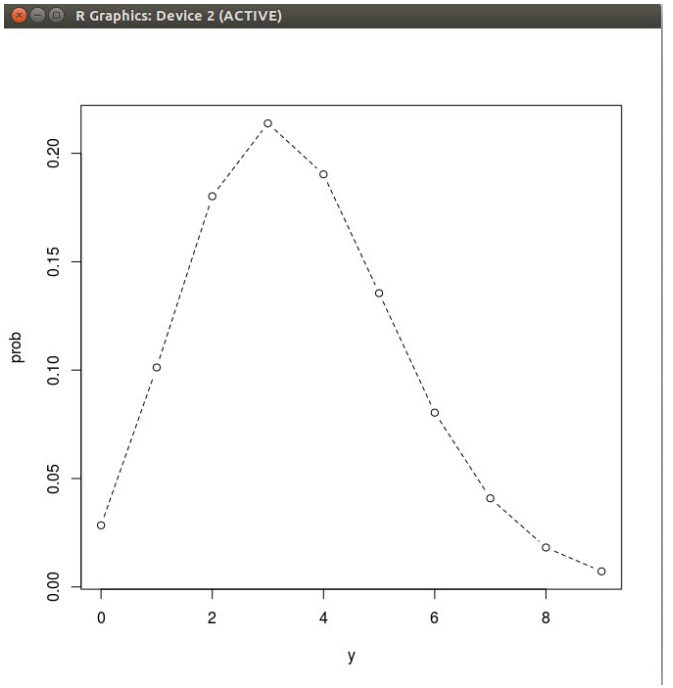
```
> prob
 [1] 0.02843882 0.10124222 0.18021114 0.21385056 0.19032700 0.13551282
 [7] 0.08040427 0.04089132 0.01819664 0.00719778
>
> plot(y,prob,type="b",lty=2)
>
```



例題データとポアソン分布

- 種子数 y_iは非負
- 下限は0, 上限は不明
- 平均と分散はおなじくらい





最尤推定

- 尤度が最大となるパラメータを探す。
- 尤度:

  ある λ におけるすべての個体 i についての確率$P(y_i|\lambda)$の積

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

$$\downarrow$$

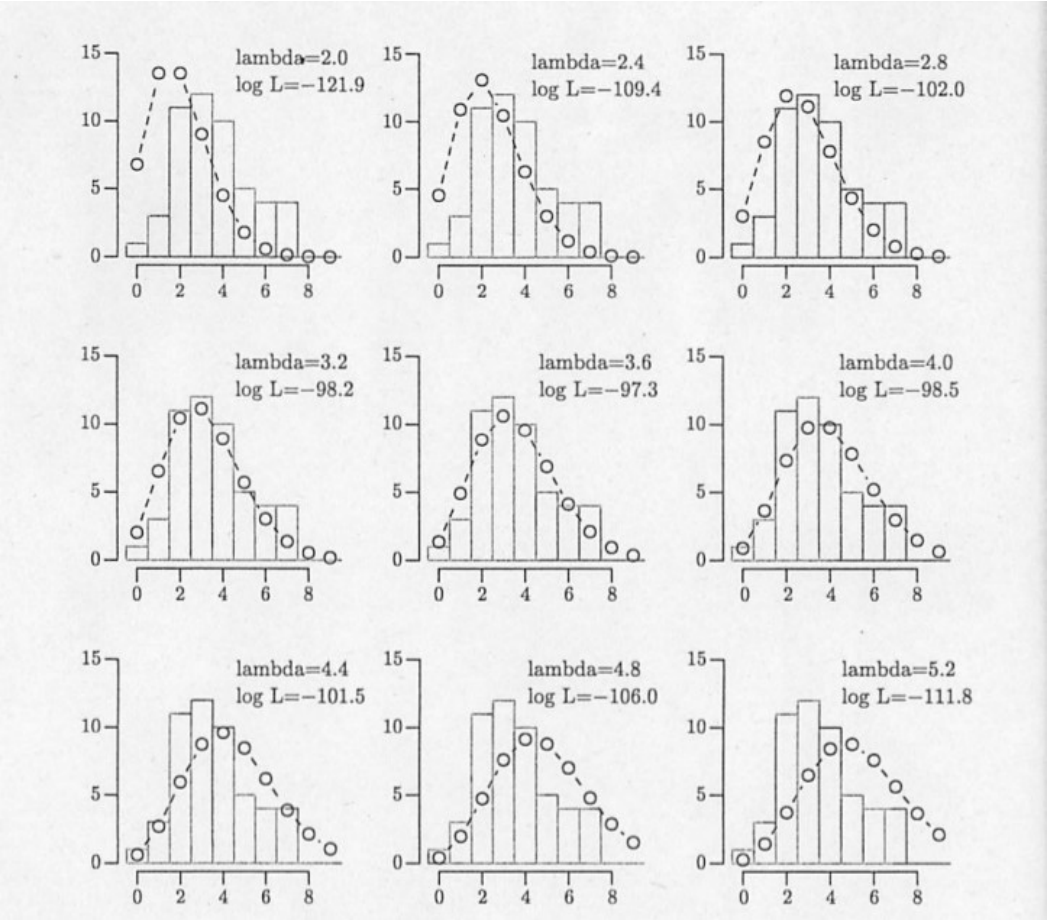
$$\log L(\lambda)=\log\prod_i \frac{\lambda^{y_i}\mathrm{e}^{-\lambda}}{y_i!}$$

$$=\sum_i \log\frac{\lambda^{y_i}\mathrm{e}^{-\lambda}}{y_i!}$$

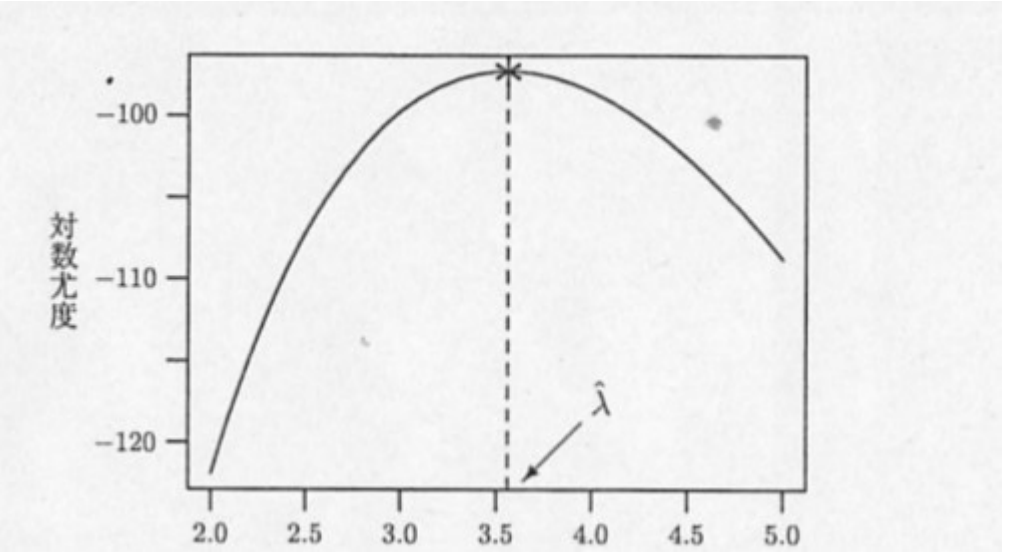
$$=\sum_i \left(y_i\log\lambda-\lambda\log e-\log\prod_k^{y_i} k\right)$$

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

パラメータ別のポアソン分布



対数尤度とλ



- 対数尤度のλについての偏微分が0となるλ

$$\log L(\lambda)=\sum_i (y_i \log\lambda-\lambda-\sum_k \log k)$$

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

$$=\frac{1}{\lambda}\sum_i y_i-50$$

$$\lambda=\frac{1}{50}\sum_i y_i \left(=3.56\right)$$

一般化

- パラメータ θ
- 尤度

$$L(\theta|Y)=\prod_i p(y_i|\theta)$$

- 対数尤度

$$\log L(\theta|Y)=\sum_i \log p(y_i|\theta)$$