

# 統計的モデリング基礎④ ～最尤推定（続き）～

鹿島久嗣  
(情報学科 計算機科学コース)

# 最尤推定

# パラメータの推定問題：

サイコロの各目の出る確率を実際の出目から推定する

- 母集団は離散分布に従うとする

- $P(X = k) = f(k)$  (ただし  $\sum_{k \in \mathcal{X}} f(k) = 1, f(k) \geq 0$ )

- たとえば（厳密な）サイコロであれば  $P(X = k) = \frac{1}{6} \approx 0.17$

- 標本抽出：

- サイコロを20回（独立に）振ったところ、

6 3 5 1 3 1 4 1 2 2 6 1 2 2 5 4 4 4 6 5 が出た

出目	1	2	3	4	5	6
回数	4	4	2	4	3	3

- 母集団のパラメータ（それぞれの目の出る確率）を推定したい

# 最尤推定とは：

標本をもっともよく再現するパラメータを推定値とする

- $n$ 個のデータ：  $x_1, x_2, \dots, x_n$  が生成される確率（尤度）：

$$L = P(X = x_1)P(X = x_2) \cdots P(X = x_n) = \prod_{i=1}^n P(X = x_i)$$

独立性を仮定しているので積になる

- サイコロの例：

– 目  $k$  が出る確率を  $p_k$ , 目  $k$  が出た回数を  $n_k$  とする

– 尤度  $L(p_1, p_2, \dots, p_n) = p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6} = \prod_{k=1}^6 p_k^{n_k}$

– これを最大化する  $p_1, p_2, \dots, p_n$  を求める（最大化問題を解く）と

$$\hat{p}_k = \frac{n_k}{n_1 + n_2 + \cdots + n_6}$$

# サイコロ（離散分布）の最尤推定： ラグランジュの未定乗数法によって推定値が求まる

- 尤度の代わりに対数尤度を最大化すると扱いやすい（解は変わらない）：

$$\log L(p_1, p_2, \dots, p_n) = \sum_{k=1}^6 n_k \log p_k$$

- 確率分布の制約:  $\sum_{k=1}^6 p_k = 1, p_k > 0$

- ラグランジュ未定乗数法：

$\{p_k\}_{k=1}^6, \lambda$  について最大化する

$$G(\{p_k\}_{k=1}^6, \lambda) = \sum_{k=1}^6 n_k \log p_k + \lambda \left( 1 - \sum_{k=1}^6 p_k \right)$$

# ベイズ決定

# 応用問題：

## どちらのサイコロが使われた？

- 2つの（いびつな）サイコロA, Bがある

ーサイコロAを20回振ったところ：

出目	1	2	3	4	5	6
回数	5	1	4	2	4	4

ーサイコロBを16回振ったところ：

出目	1	2	3	4	5	6
回数	2	8	2	2	1	1

# 応用問題：

## どちらのサイコロが使われた？

- (いびつな) サイコロA, Bのパラメータの最尤推定値は：

ーサイコロA：

出目	1	2	3	4	5	6
確率	5/20	1/20	4/20	2/20	4/20	4/20

ーサイコロB：

出目	1	2	3	4	5	6
確率	2/16	8/16	2/16	2/16	1/16	1/16



# 応用問題：

## どちらのサイコロが使われた？

- (いびつな) サイコロA, Bのパラメータの最尤推定値は：

ーサイコロA：

出目	1	2	3	4	5	6
確率	5/20	1/20	4/20	2/20	4/20	4/20

ーサイコロB：

出目	1	2	3	4	5	6
確率	2/16	8/16	2/16	2/16	1/16	1/16


- 今、2つのサイコロのいずれかを選んで (Cとする) 5回振ったところ：

出目	1	2	3	4	5	6
回数	1	1	0	2	0	1

- 使われたサイコロはA, Bのいずれだろうか？ (C=A or C=B?)

# ベイズ決定： 事後確率によって決定する

- A, B どちらのサイコロを選んだかを確率変数  $X$  で表す
  - 事前確率：でたらめに選ぶと  $P(X = A) = P(X = B) = 1/2$
  - 何も情報がなければこれ以上はわからない
- 事後分布：C (A, Bのいずれか) を振って出たデータ  $\mathcal{D}$  を見たあとの、 $X$  の確率分布  $P(X|\mathcal{D})$
- ベイズ決定：事後確率が  $P(X = A|\mathcal{D}) > P(X = B|\mathcal{D})$  であれば、A が使われた可能性が高いと判断できる
- 事後確率の計算：  $P(X|\mathcal{D}) = \frac{P(\mathcal{D}|X)P(X)}{P(\mathcal{D})}$  (ベイズの定理)



出目	1	2	3	4	5	6
回数	1	1	0	2	0	1

# 事後確率の計算：

## ベイズの定理と最尤推定で事後確率を計算

- 事後確率の計算には「ベイズの定理」をつかう：

$$P(X|\mathcal{D}) = \frac{P(\mathcal{D}|X)P(X)}{P(\mathcal{D})}$$

$P(X)$ は事前確率  
(今回は1/2)

*T. Bayes.*



– 判断基準： $P(X = A|\mathcal{D}) \geq P(X = B|\mathcal{D})$

$$\Leftrightarrow P(\mathcal{D}|X = A)P(X = A) \geq P(\mathcal{D}|X = B)P(X = B)$$

– 注意：分母 $P(\mathcal{D}) = \sum_X P(\mathcal{D}|X)P(X)$ は今回は計算する必要はない

- サイコロのパラメータ $\{p_k^A\}_{k=1}^6, \{p_k^B\}_{k=1}^6$ は最尤推定によって推定

サイコロCの出目回数

- $P(\mathcal{D}|X = A) = \prod_{k=1}^6 p_k^A n_k^C \geq P(\mathcal{D}|X = B) = \prod_{k=1}^6 p_k^B n_k^C$ で判断

# 線形回帰モデルの確率的解釈

# 最尤推定：

データをもっともよく再現するパラメータを推定値とする

- $n$ 個のデータ  $x^{(1)}, x^{(2)}, \dots, x^{(n)}$  から確率モデル  $f_{\theta}(x)$  のパラメータ  $\theta$  を推定したい

- $n$ 個のデータが（互いに独立に）生成される確率（尤度）：

$$L(\theta) = \prod_{i=1}^n f_{\theta}(x^{(i)})$$

- 尤度最大になるパラメータを推定値  $\hat{\theta}$  とする

実際には対数尤度で扱うことが多い

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i=1}^n f_{\theta}(x^{(i)}) = \operatorname{argmax}_{\theta} \sum_{i=1}^n \log f_{\theta}(x^{(i)})$$

— もっともデータを生成する確率が高い（「最も尤もらしい」）

# 線形回帰モデルの最尤推定： 線形回帰の確率モデルを考える

- データ：  $\mathbf{x} = (x^{(1)}, x^{(2)}, \dots, x^{(n)})$  と  $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(n)})$  に  
線形モデル：  $g(x) = \beta x + \alpha$  を当てはめる
- 最小二乗法：  $\ell(\alpha, \beta) = \sum_{i=1}^n \left( y^{(i)} - (\beta x^{(i)} + \alpha) \right)^2$  を最小化
- 一方、線形回帰モデルに対応する確率モデルを仮定する：
  - 正規分布：  $y^{(i)}$  は平均  $\beta x^{(i)} + \alpha$  , 分散  $\sigma^2$  の正規分布に従う
  - 確率密度：  $f(y^{(i)} | x^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - (\beta x^{(i)} + \alpha))^2}{2\sigma^2}\right)$
  - 「平均的に」回帰直線  $y = \beta x + \alpha$  に乗るデータを生成するモデル

# 線形回帰モデルの最尤推定：

## 線形回帰の確率モデルの最尤推定 = 最小二乗法

- 線形回帰モデルに対応する確率モデルを考える：

- 確率密度関数：
$$f(y^{(i)} | x^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - (\beta x^{(i)} + \alpha))^2}{2\sigma^2}\right)$$

- 対数尤度：
$$L(\alpha, \beta) = \sum_{i=1}^n \log f(y^{(i)} | x^{(i)})$$

注：ここでは $y^{(i)}$ が「再現されるべき」データ

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^n (y^{(i)} - (\beta x^{(i)} + \alpha))^2 + \text{const.}$$

- 対数尤度を $\alpha, \beta$ について最大化すること（最尤推定）  
= 二乗誤差を $\alpha, \beta$ について最小化すること（最小二乗法）

# 線形回帰モデルの最尤推定： 分散の最尤推定量

- 確率密度関数： $f(y^{(i)} | x^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - (\beta x^{(i)} + \alpha))^2}{2\sigma^2}\right)$

- 分散については、対数尤度：

$$L(\sigma^2) = n \log \frac{1}{\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^{(i)} - (\beta x^{(i)} + \alpha))^2 + \text{const.}$$

- $L(\sigma^2)$ を最大化する最尤推定量は：

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - (\beta x^{(i)} + \alpha))^2$$

※ 以上の議論は重回帰モデルの場合も同様



# 最尤推定の性質

# 最尤推定量の性質： 一貫性

- パラメータ $\theta$ の推定量として $\hat{\theta}$ を得たとする（例えば最尤推定で）
- 推定量の良さはどのように評価するか？

– 不偏性  $E[\hat{\theta}] = \theta$  : 推定量の期待値が真の値に一致する

- $E$ は様々な標本の採り方についての期待値を表す
- たとえば、平均の最尤推定量は不偏性をもつが、分散の最尤推定量はもたない

– 一貫性：標本サイズを大きくしていくと真の値に一致する：

$$\hat{\theta} \xrightarrow{n \rightarrow \infty} \theta$$

- 最尤推定は、適当な条件のもと一貫性をもつ

# 漸近正規性：

## 最尤推定量は漸近的に正規分布に従う

- 最尤推定量の分布は  $n \rightarrow \infty$  で、真のパラメータ  $\theta$  を平均とする正規分布に従う

- もう少し厳密にいうと：

$\sqrt{n}(\hat{\theta} - \theta)$  の分布が平均0、分散  $I(\theta)^{-1}$  の正規分布に近づく

- $I(\theta)$  はフィッシャー情報量：

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right]$$
$$= - \int \left( \frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right) f(x|\theta) dx$$

$I(\theta)$  が大きいほど、対数尤度関数が、真値の周りで「尖っている」イメージ

- $n \rightarrow \infty$  で  $\hat{\theta} \rightarrow \theta$

# ポアソン回帰

# 最尤推定：

データをもっともよく再現するパラメータを推定値とする

- $n$ 個のデータ  $x^{(1)}, x^{(2)}, \dots, x^{(n)}$  から確率モデル  $f_{\theta}(x)$  のパラメータ  $\theta$  を推定したい

- $n$ 個のデータが（互いに独立に）生成される確率（尤度）：

$$L(\theta) = \prod_{i=1}^n f_{\theta}(x^{(i)})$$

実際には対数  
尤度で扱うこと  
が多い

- 尤度最大になるパラメータを推定値  $\hat{\theta}$  とする

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i=1}^n f_{\theta}(x^{(i)}) = \operatorname{argmax}_{\theta} \sum_{i=1}^n \log f_{\theta}(x^{(i)})$$

—もっともデータを生成する確率が高い（「最も尤もらしい」）

# ポアソン分布の最尤推定： 標本平均がパラメータの最尤推定量になる

- ポアソン分布： $f_{\lambda}(Y = y) = \frac{\lambda^y}{y!} \exp(-\lambda)$

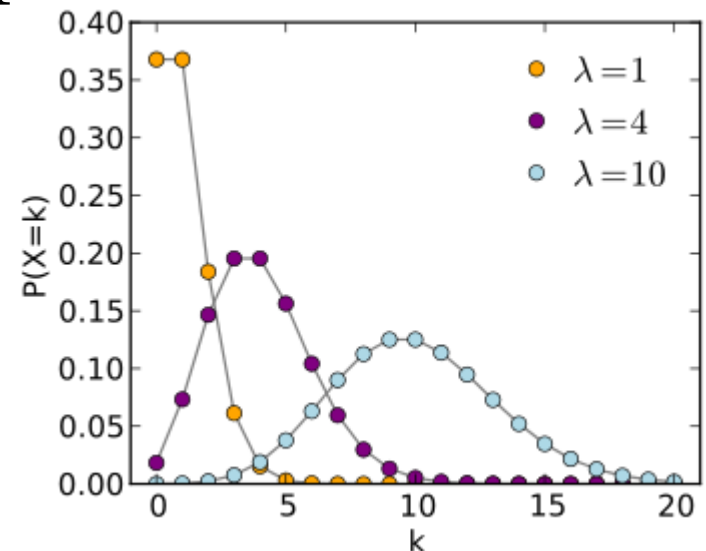
$\lambda > 0$ は平均に相当するパラメータ

- データ： $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ に対する対数尤度：

$$L(\lambda) = \sum_{i=1}^n \log f_{\lambda}(Y = y^{(i)}) = \log \lambda \sum_{i=1}^n y^{(i)} - n\lambda + \text{const.}$$

- パラメータの最尤推定量：

$$\hat{\lambda} = \frac{\sum_{i=1}^n y^{(i)}}{n}$$



[https://en.wikipedia.org/wiki/Poisson\\_distribution#/media/File:Poisson\\_pmf.svg](https://en.wikipedia.org/wiki/Poisson_distribution#/media/File:Poisson_pmf.svg)

# ポアソン回帰： 非負整数の回帰モデル

- 例えば、ある機械の各日の故障件数をモデル化したいとする
  - 曜日や気温などに依存して平均的な故障件数が変わるとする
- 独立変数（曜日など）に依存する回数のモデル：ポアソン回帰

$$f_{\boldsymbol{\beta}}(Y = y) = \frac{(\exp(\boldsymbol{\beta}^T \mathbf{x}))^y}{y!} \exp(-\exp(\boldsymbol{\beta}^T \mathbf{x}))$$

–ポアソン分布の平均が線形モデルで表されるとする：

- ポアソン分布： $f_{\lambda}(Y = y) = \frac{\lambda^y}{y!} \exp(-\lambda)$
  - 重回帰モデル： $\lambda = \exp(\boldsymbol{\beta}^T \mathbf{x})$
- } 組み合わせる

# ポアソン回帰の最尤推定： 解析解は得られなさそう...

- 独立変数： $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)})$  #  $n$ 日分の測定
- 従属変数： $(y^{(1)}, y^{(2)}, \dots, y^{(n)})$  #  $n$ 日分の故障数
- 対数尤度（最大化問題）：

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^n \log \frac{\left( \exp(\boldsymbol{\beta}^\top \mathbf{x}^{(i)}) \right)^{y^{(i)}}}{y^{(i)}!} \exp(-\exp(\boldsymbol{\beta}^\top \mathbf{x}^{(i)})) \\ &= \sum_{i=1}^n y^{(i)} \boldsymbol{\beta}^\top \mathbf{x}^{(i)} - \sum_{i=1}^n \exp(\boldsymbol{\beta}^\top \mathbf{x}^{(i)}) + \text{const.} \end{aligned}$$

- これを最大化する $\boldsymbol{\beta}$ を求めたいが、解析解は得られない



# 最尤推定の利点： モデリングの自動化

- 最尤推定の利点：確率モデルの形（データの生成プロセスの仮定）を決めればモデルパラメータが自動的に決まる
  - ーただし、最後に最大化問題を解いて、パラメータ推定量を求める必要がある
  - ー離散分布、ポアソン分布、正規分布などは解析的に解が求まる
  - ー線形回帰（正規分布でノイズが載る）は連立方程式（いちおう解析的な解）
  - ーただし、他の多くのモデルでは、最適化問題を数値的に解く必要がある