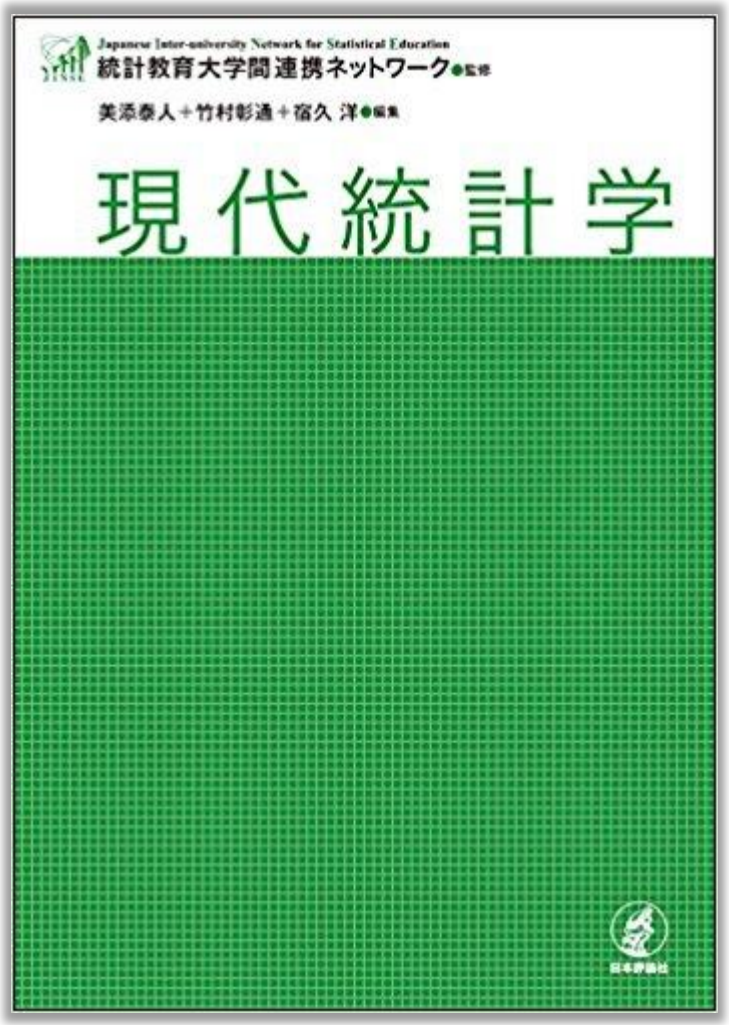
KYOTO UNIVERSITY

統計的モデリング基礎④
～最尤推定～

鹿島久嗣
（情報学科 計算機科学コース）

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

（いろいろな話題についての）参考書



現代統計学

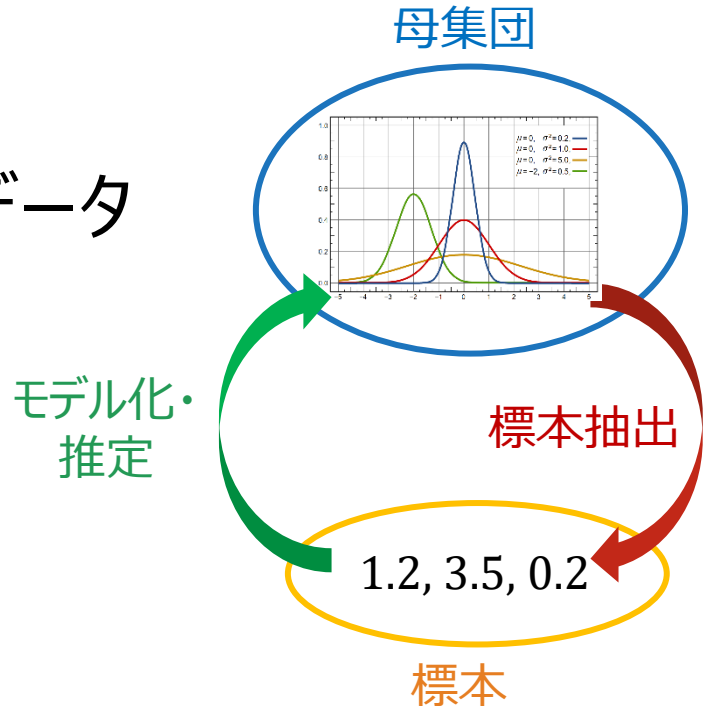
出版社：日本評論社
発刊年月： 2017.03
ISBN ： 978-4-535-78818-3
A5判；256ページ

幅広いトピックで基本的事項がコンパクトにまとまっている

最尤推定

統計モデリングの考え方：部分から全体について知る

- 母集団：確率分布で表される、我々が本当に興味のある集合
  - 分布のクラスやパラメータで指定されるとする
- 標本：実際に観測できる母集団の一部
  - 確率分布に従って抽出された具体的なデータ
- 目的：
  標本から母集団について推測する
  （標本抽出の逆）
  - パラメータを推定する（どうやって？）



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

サイコロのパラメータ推定問題へのひとつの解：出た目の回数の割合で推定する

- ひとつのアイディア：

  20回中で1が4回出たのだから$P(X=1) \approx \frac{4}{20} = 0.2$と推定する

出目	1	2	3	4	5	6
回数	4	4	2	4	3	3
確率の推定値	0.2	0.2	0.1	0.2	0.15	0.15

- 正解が約0.17なので悪くない…
- この推定値はどのような原理に基づいているのか？

最尤推定：
確率分布の代表的な推定手法のひとつ

- 標本からの母集団確率分布の推定
- 代表的な推定手法
  - 最尤推定
  - モーメント推定
  - ベイズ推定
  - …

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： $x_1, x_2, \ldots, x_n$ 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$L(p_1, p_2, \ldots, p_n) = p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6} = \prod_{k=1}^{6} p_k^{n_k}$
  - これを最大化する$p_1, p_2, \ldots, p_n$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）：

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

- 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：

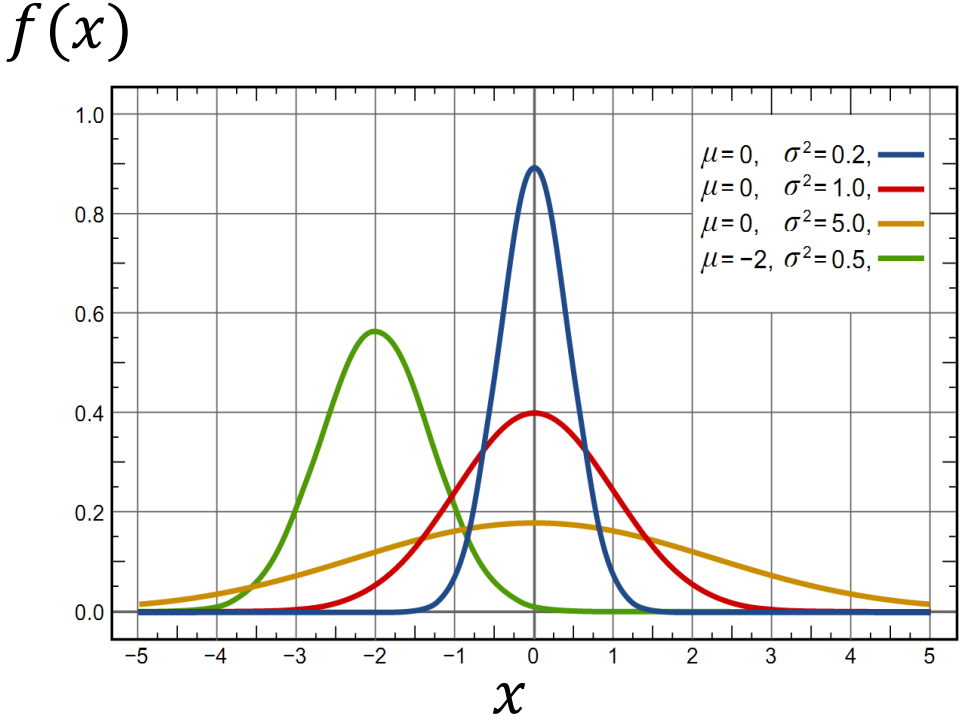
$\{p_k\}_{k=1}^{6}, \lambda$ について最大化する

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

練習：
正規分布のパラメータの最尤推定

- 正規分布：$f(x) = N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- パラメータ：平均μと分散σ^2の最尤推定量を求めてみよう

1. データ：$x_1, x_2, \ldots, x_n$ に対する対数尤度をつくる
2. パラメータについての最大化問題を解く



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
  - 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

  - 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：
  - 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

  - 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：
  - 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

  - 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 $P(X = \mathrm{A}) = P(X = \mathrm{B}) = 1/2$
  - 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 $P(X|\mathcal{D})$

出目	1	2	3	4	5	6
回数	1	1	0	2	0	1

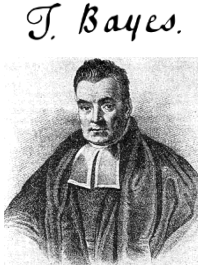
- ベイズ決定：事後確率が$P(X = \mathrm{A}|\mathcal{D}) > P(X = \mathrm{B}|\mathcal{D})$であれば、Aが使われた可能性が高いと判断できる
- 事後確率の計算： $P(X|\mathcal{D}) = \frac{P(\mathcal{D}|X)P(X)}{P(\mathcal{D})}$（ベイズの定理）

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）



  - 判断基準：$P(X = \mathrm{A}|\mathcal{D}) \gtrless P(X = \mathrm{B}|\mathcal{D})$
    $\leftrightarrow\ P(\mathcal{D}|X = \mathrm{A})P(X = \mathrm{A}) \gtrless P(\mathcal{D}|X = \mathrm{B})P(X = \mathrm{B})$
    - 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$\{p_k^{\mathrm{A}}\}_{k=1}^6, \{p_k^{\mathrm{B}}\}_{k=1}^6$は最尤推定によって推定

サイコロCの出目回数

- $P(\mathcal{D}|X = \mathrm{A}) = \prod_{k=1}^6 {p_k^{\mathrm{A}}}^{n_k^{\mathrm{C}}} \gtrless P(\mathcal{D}|X = \mathrm{B}) = \prod_{k=1}^6 {p_k^{\mathrm{B}}}^{n_k^{\mathrm{C}}}$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 $x_1, x_2, \ldots, x_n$ から確率モデル$f(x \mid \theta)$のパラメータθ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）：

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- 尤度最大になるパラメータを推定値$\hat{\theta}$とする

$$\hat{\theta} = \mathrm{argmax}_\theta \prod_{i=1}^{n} f(x_i \mid \theta) = \mathrm{argmax}_\theta \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

実際には対数尤度で扱うことが多い

–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：$\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ と $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right)$ に線形モデル： $g(x) = \beta x + \alpha$ を当てはめる

- 最小二乗法：$\ell(\alpha, \beta) = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta x^{(i)} + \alpha\right)\right)^2$ 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 σ^2 の正規分布に従う
  - 確率密度：$f\left(y^{(i)} \mid x^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - (\beta x^{(i)} + \alpha)\right)^2}{2\sigma^2}\right)$
  - 「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：$f\left(y^{(i)} \mid x^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)}-(\beta x^{(i)}+\alpha)\right)^2}{2\sigma^2}\right)$

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

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

- 対数尤度をα, βについて最大化すること（最尤推定）
  ＝二乗誤差をα, β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：$f\left(y^{(i)} \mid x^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)}-(\beta x^{(i)}+\alpha)\right)^2}{2\sigma^2}\right)$

- 分散については、対数尤度：

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

- $L(\sigma^2)$を最大化する最尤推定量は：

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

※ 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：一致性

- パラメータθ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
  - 一致性：標本サイズを大きくしていくと真の値に一致する：

$$\hat{\theta} \xrightarrow[n \to \infty]{} \theta$$

- 最尤推定は、適当な条件のもと一致性をもつ

漸近正規性：
最尤推定は漸近正規性をもつ

- 最尤推定量の分布は$n \to \infty$で、真のパラメータθ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：

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

$$= -\int\left(\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right) f(x|\theta)\mathrm{d}x$$

$I(\theta)$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

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

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
  - 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
  - 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
  - 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）
  - 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

ポアソン回帰

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：$P(Y = y \mid \lambda) = \frac{\lambda^y}{y!}\exp(-\lambda)$

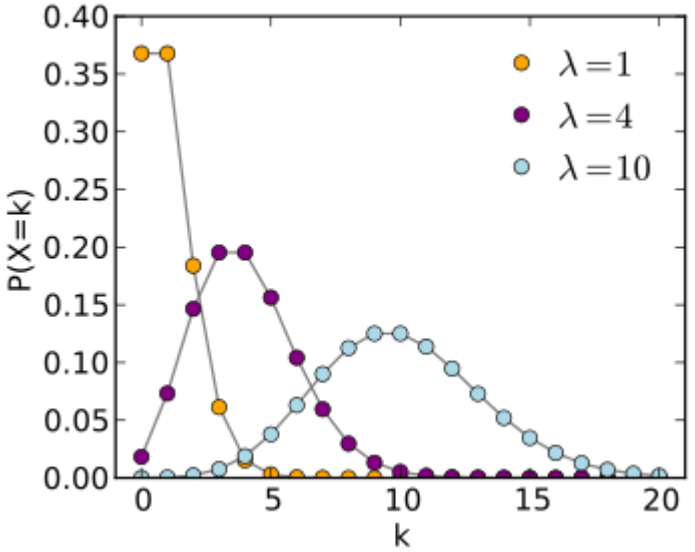
$\lambda > 0$は平均に相当するパラメータ

- データ：$y_1, y_2, \ldots, y_n$ に対する対数尤度：

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

- 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

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

$$P(Y = y \mid \mathbf{x}, \boldsymbol{\beta}) = \frac{\left(\exp(\boldsymbol{\beta}^\top \mathbf{x})\right)^y}{y!} \exp(-\exp(\boldsymbol{\beta}^\top \mathbf{x}))$$

  - 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：$P(Y = y \mid \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$
    - 重回帰モデル：$\lambda = \exp(\boldsymbol{\beta}^\top \mathbf{x})$

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

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \frac{\left(\exp\left(\boldsymbol{\beta}^{\top}\mathbf{x}^{(i)}\right)\right)^{y^{(i)}}}{y^{(i)}!} \exp\left(-\exp\left(\boldsymbol{\beta}^{\top}\mathbf{x}^{(i)}\right)\right)$$

$$= \sum_{i=1}^{n} y^{(i)}\boldsymbol{\beta}^{\top}\mathbf{x}^{(i)} - \sum_{i=1}^{n} \exp\left(\boldsymbol{\beta}^{\top}\mathbf{x}^{(i)}\right) + \text{const.}$$

- 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