KYOTO UNIVERSITY

統計的モデリング基礎Ⓧ
～さまざまな確率モデル～
（生存期間のモデル）

鹿島久嗣
（情報学科 計算機科学コース）

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

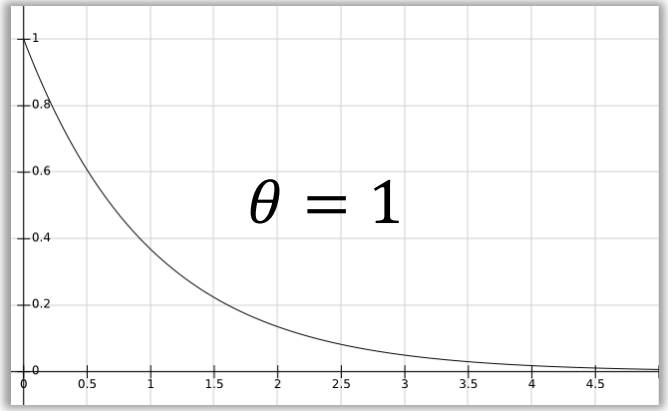
生存期間のモデル：期間を確率変数とするモデル

- 期間（非負の実数）を確率変数とするようなモデル：
  - 商品の寿命、患者の生存期間、…
  - 一方、ポアソン分布は回数（非零の整数）のモデル
- 生存期間の確率変数T：$\Pr(T \leq t) = F(t) = \int_0^t f(\tau)\,\mathrm{d}\tau$
  - 確率密度関数$f(t)$：時刻tまで生存していて、時刻$t + \Delta t$までの間の死亡確率が$f(t)\Delta t$（「時刻tまで生存」かつ「$t \sim t + \Delta t$で死亡」）
    - $f(t) = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t}$
  - $\Pr(T > t) = S(t) = 1 - F(t)$：時刻$t$以降も生存する確率
    （少なくとも時刻tまでは生存する）

生存関数

指数分布モデル：
もっとも単純な生存期間のモデル

- $f(t)$ ：時刻tまで生存していて、時刻$t + \Delta t$までの間の死亡確率が$f(t)\Delta t$（「時刻tまで生存」かつ「$t \sim t + \Delta t$ で死亡」）
- 指数分布モデル：$f(t) = \theta \exp(-\theta t)$
  - $\theta > 0$ ：モデルパラメータ



- 生存期間T：

$$\Pr(T \leq t) = F(t) = \int_0^t f(\tau)\,\mathrm{d}\tau = 1 - \exp(-\theta t)$$

  - $\mathrm{E}[T] = \frac{1}{\theta}, \mathrm{Var}[T] = \frac{1}{\theta^2}$

- 独立変数によってパラメータが変わる場合：$\theta = \exp(\boldsymbol{\beta}^\top \mathbf{x})$

ハザード関数：
ある時刻の死亡リスクを表す関数

- $f(t)$ は「時刻tまで生存している」かつ「次の瞬間に死亡する」可能性を表す（ちょっと解釈しにくい）
- 瞬間瞬間の死亡リスクをみたほうがわかりやすい？
  - 「時刻tまで生存している」という条件のもとでの「次の瞬間に死亡する」可能性（条件付確率）をみる
- ハザード関数： $h(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}$

$S(t) = 1 - F(t)$ ：
生存関数（少なくとも時刻tまでは生存する確率）

- ハザード関数の時間変化：
  $\frac{\mathrm{d}h(t)}{\mathrm{d}t} > 0$ のとき、リスクが時間とともに増加（< 0 であれば減少）

ワイブル分布：
指数分布の一般化

- 指数分布モデルはリスクが時間に関わらず一定
  - 指数分布のハザード関数： $h(t) = \frac{f(t)}{S(t)} = \frac{\theta \exp(-\theta t)}{\exp(-\theta t)} = \theta$ (定数)
- ワイブル分布モデル：

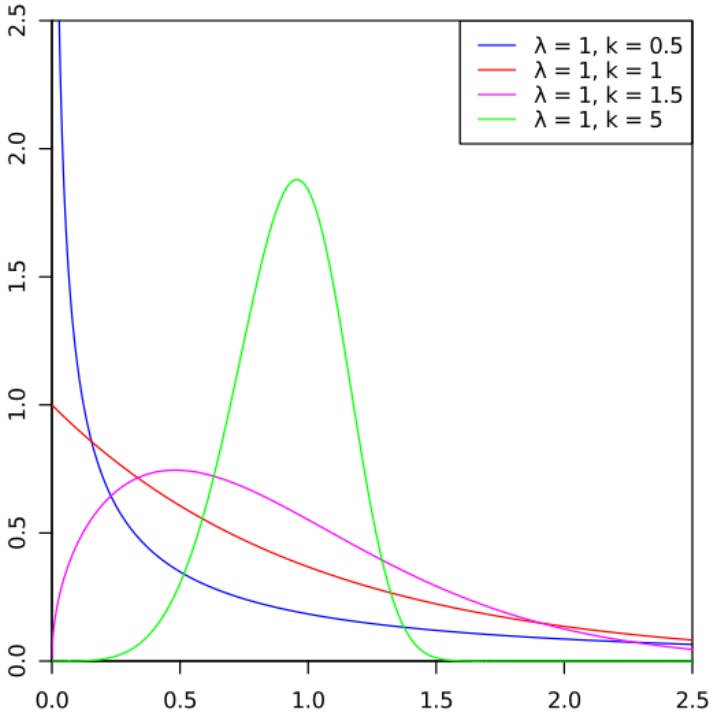
$$f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{t}{\lambda}\right)^{k}\right\}, k, \lambda > 0$$

  - $k = 1$のとき指数分布（ $\theta = 1/\lambda$）

$$f(t) = \theta \exp(-\theta t)$$

- 独立変数を取り込む場合：

$$\lambda = \exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x})$$



https://en.wikipedia.org/wiki/Weibull_distribution#/media/File:Weibull_PDF.svg

ワイブル分布：
パラメータによってハザード関数の時間的増減が決まる

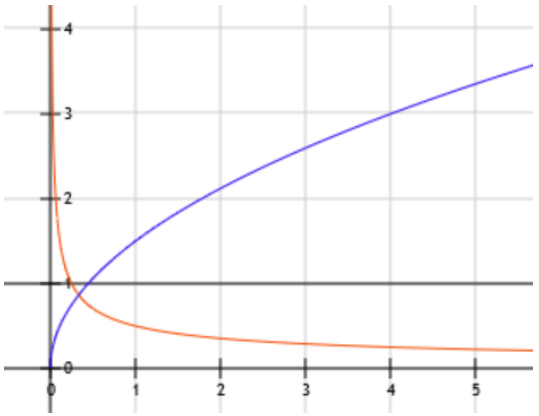
- ワイブル分布の生存関数：

$$S(t) = \int_t^{\infty} \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{t}{\lambda}\right)^{k}\right\} \mathrm{d}t = \exp\left\{-\left(\frac{t}{\lambda}\right)^{k}\right\}$$

- ハザード関数： $h(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}$

  - $k = 1$ のとき $h(t) = 1/\lambda$
  - $k > 1$ のとき $\frac{\mathrm{d}h(t)}{\mathrm{d}t} > 0$
  - $k < 1$ のとき $\frac{\mathrm{d}h(t)}{\mathrm{d}t} < 0$

kによって決まる



生存時間モデルの最尤推定：生存期間の確率密度関数$f(t)$を最尤推定

- データ $\{t^{(1)}, t^{(2)}, \ldots, t^{(N)}\}$ ：
  - N個の独立な観測（生存期間がちょうど$t^{(i)}$）
- 尤度関数$L(\theta) = \prod_{i=1}^{N} f\left(t^{(i)}\right)$

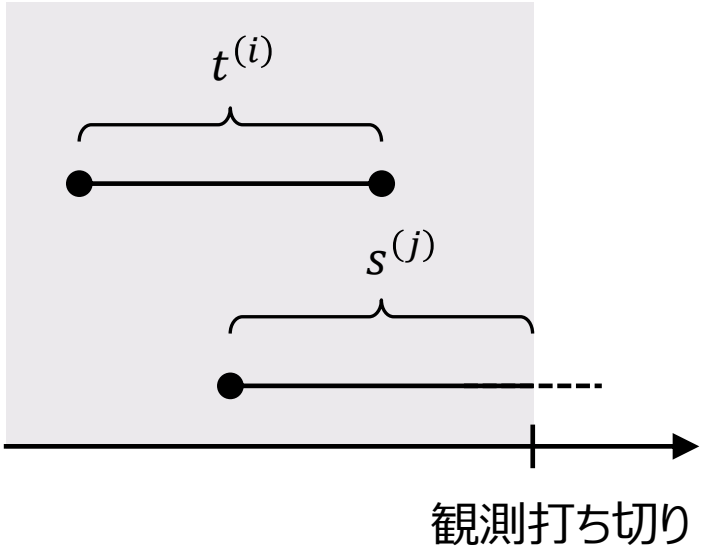
言い換えれば、$t^{(i)}$まで生きていて次の瞬間死亡したという観測データ

  - 確率密度関数$f(t)$ ：時刻tまで生存していて、時刻$t + \Delta t$までの間の死亡確率が$f(t)\Delta t$
  - 指数分布モデルの場合：$L(\theta) = \prod_{i=1}^{N} \theta \exp\left(-\theta t^{(i)}\right) \Delta t$
    - 対数尤度にすると $\log L(\theta) = N \log \theta - \theta \sum_{i=1}^{N} t^{(i)}$
    - 最尤推定量は $\hat{\theta} = \frac{N}{\sum_{i=1}^{N} t^{(i)}}$

打ち切りがある場合の最尤推定：打ち切りデータに対して生存関数$S(t)$を当てはめる

- データ$\{t^{(1)}, t^{(2)}, \ldots, t^{(N)}\} \cup \{s^{(1)}, s^{(2)}, \ldots, s^{(M)}\}$：
  - N個の生存期間データに加えて、M個の打ち切りデータ（少なくとも$s^{(i)}$期間生存）
- 尤度関数：

$$L(\theta) = \prod_{i=1}^{N} f\left(t^{(i)}\right) \cdot \prod_{j=1}^{M} S\left(t^{(j)}\right)$$



  - 生存関数$S(t) = \int_t^{\infty} f(t)\mathrm{d}t$：少なくとも$t$期間は生存している確率
  - 指数分布の場合：$S(t) = \exp(-\theta t)$
    - $\log L(\theta) = N \log \theta - \theta\left(\sum_{i=1}^{N} t^{(i)} + \sum_{j=1}^{M} s^{(j)}\right)$

今回の話題：
生存期間のモデル

- 生存期間のモデル
  - ハザード関数
  - 生存期間モデルの最尤推定