Chapter 11. Open Population Models

2010/12/19 (Sun.) 飯島勇人*†

飯島の担当部分の概要

1. 開放系（個体が研究対象の個体群を出入りする）の動物データのモデリング
   - 打ち切り（左、中間、右）がある生残データ
   - 多項分布を用いた標識再捕獲データのモデリング
2. 空間的な出入りは扱わない

訳語

- 本文の"": 「」で表記
- lifetime distribution: 寿命分布
- hazard function: ハザード関数
- cell probability: 分割確率（多項分布のパラメータ）
- force of mortality: 「死に追いやる力」

始め〜11.1 節まで

この章では、研究の間に動物が対象とする個体群を出入りできる開放された個体群に関するモデルを考える（9・10 章では閉鎖個体群を扱った）。まず生残に関するモデルから始め、生残及び個体数両方に関する標識再捕獲モデルへと進む（p.239 第 1 パラグラフ）。

（p.239 第 2 パラグラフは省略）

（p.240 第 1 パラグラフは省略）

個体群を個体が出入りするケースとして、出生と死亡が挙げられる。出生と死亡が調査期間の始まりと終わりに対応することはまず無いので、以下の 2 種類の打ち切りが発生する。

- 左打ち切り：調査開始前から存在する
- 右打ち切り：調査終了後も生残している

まず、このような開放個体群のモデリングについて議論する（p.240 第 2 パラグラフ）。

* 山梨県森林総合研究所森林保護科研究員（注！（独）森林総合研究所とは一切関係ありません）
† 連絡先: hayato.iijima@gmail.com または http://www7.atwiki.jp/hayatoiijima/

11.1 CONTINUOUS-TIME SURVIVAL MODELS

（個体群動態のモデリングに）共通する統計的な問題として、調査開始時点 t_0 における、サンプリングした全ての生残個体に寿命分布を当てはめる問題がある。そのようなデータを使うモデルは「生存モデル」と呼ばれる。理想的には、サンプリングした n 個の個体について寿命 $T_1, T_2, ..., T_n$ を観察する。これらは、$f(t)$ で表される確率密度関数を持った適切な寿命分布を仮定してモデル化される。寿命分布はしばしば、$f(t)$ ではなく $h(t)$ と $f(t)$ の関係を示した以下のような「ハザード関数 $h(t)$」で定義される。

$$h(t) = \frac{f(t)}{1 - F(t)}$$

ここで $F(t)$ は累積分布関数 $F(t) = \int_0^t f(s)ds$ である。ハザード率 $h(t)$ は、ある時点 t までは生残し、ある時点 t において失敗または死亡する瞬時的なリスクと解釈できる（p.240 第 3 パラグラフ）。

ハザード関数については様々な物が開発されている（p.240 第 4 パラグラフ）。

- 指数生残モデル：$f(t) = \lambda e^{-\lambda t}$ かつ全ての t において $h(t) \equiv \lambda$ と一定である。最も単純なモデル。
- ワイブルモデル：単調増加または単調減少を記述するモデル。
- Frailty モデル：ハザード率に個体固有の変量効果を割り当てる
- Cox 比例ハザードモデル：現在最も一般的に用いられる生残モデル。以下で詳細に解説。

コックス比例ハザードモデルでは、個体 i のハザード関数は、一連の共変量 z_i とパラメータのベクタ β で以下のように記述される。

$$h_i(t) = \lambda_0(t) e^{z_i' \beta}$$

$\lambda_0(t)$ は時間的に連続な関数であり、ある個体について全ての共変量が 0 と仮定した場合のハザード関数、すなわち基礎的なハザード関数と解釈できる。Cox が 1972 年に示した式においては、この基礎的なハザード関数は任意の値を取ることができ、その意味ではこのモデルはノンパラメトリックと見なせる。普通の頻度論的なやり方では、部分尤度（Cox 1975）を用いてモデルを当てはめる、つまり、興味のあるパラメータ（この場合 β）に依存する要因は保持し、重要でないパラメータに密接に関連している要因（この場合関数 $\lambda_0(t)$ のパラメータ）を捨てていく（p.241 第 1 パラグラフ）。

生存モデルは必ず打ち切りに遭遇する。前述した打ち切り以外に、中間打ち切りがある（死亡は観測と観測の間に発生するから）。しかし、ひとまず右打ち切りに集中する（p.241 第 2 パラグラフ）。

11.1.1 Right Censoring and the CDL

寿命が長い個体は、短い個体よりも打ち切りに遭遇する可能性が高い。推論に偏りを生じさせないようにするためには、この打ち切りを無視することはできない（p.241 第 3 パラグラフ）。

この右打ち切りのデータをモデリングするために、まず、完全なデータによる尤度（CDL）という観点からモデルを記述する。完全なデータとは、$T = \{T^{obs}, T^{mis}\}$（それぞれ、完全に観測された寿命データ、打ち切られた寿命データのベクタ）、C（打ち切られた回数のベクタ）、D（観測 T_i と対応した打ち切りかどうかを示す D_i のベクタ）がわかっているデータである。θ が T_i の分布を支配する一連のパラメータであるとすると、CDL は、

$$\begin{aligned}\mathcal{L}(T^{min}, \theta | C, T^{obs}, D) &\propto [T, C, D|\theta] && (11.1)\\ &= [D|T, C, \theta][T|C, \theta][C|\theta] && (11.2)\\ &\propto [D|T, C][T|\theta] && (11.3)\end{aligned}$$

- 式 11.1 から 11.2 へは、条件付き確率の定理を使って単純にした
- 式 11.2 から 11.3 へは、まず生存期間と打ち切り期間が独立であるという仮定をおくので、$[T|C, \theta] \equiv [T|\theta]$ とした
- 次に、C の分布は T の分布を支配するパラメータに影響を受けない、すなわち $[C|\theta] \equiv [C]$ と仮定した
- CDL から C^{mis} を除くので、$[C]$ は単に比例定数に吸収される: $[C]$ は T^{mis} や θ といった変数に影響しない
- D は θ に依存せず、T と C に完全に依存して決定されるという事実を利用した（p.241 第 4 パラグラフ）

式 11.3 の右辺の第一項を考えよう。D_i はそれぞれ $T_i \leq C_i$ となる事象を示すものである。T_i と C_i が与えられた元で、D は $T_i \leq C_i$ となるかどうかという 0、1 に等しい、生起確率を持つベルヌーイ分布であると考えることができる; そのため $[D_i|T_i, C_i] = B(1, \mathrm{I}(T_i \leq C_i))$ である。CDL は単に、ベルヌーイ試行の尤度と観測されたかどうかの T_i の尤度の積である（p.242 第 1 パラグラフ）。

この CDL の単純さは、BUGS コードを書かなければならなくなったときに明確になる（p.242 第 2 パラグラフ）。

対照的に、観測されたデータの尤度（ODL）に対してこのモデルを当てはめる一般的な方法は以下の式で与えられ、

$$\begin{aligned}\mathrm{ODL} &\propto [T^{obs}, d|\theta, C]\\ &= \int_{T^{mis}} [T^{mis}, T^{obs}, d|\theta, C] dT^{mis}\end{aligned}$$

ランダムな生存時間のサンプルについては以下のようになる。

$$[y^{obs}, d|\theta, C_i] = \prod_{i=1}^{n} f(y_i)^{1-d_i}(1 - F(C_I))^{d_i}$$

ここで $f(y)$ は確率密度関数で $F(y)$ は累積分布関数である（p.242 第 4 パラグラフ）。

上で与えられたモデルの計算にあたり、打ち切り時間 C_i は生存時間に無関係であるとした。多くの野生生物の研究において、打ち切りは研究者が制御できない原因の結果起こる。そのため、打ち切りの観測値はその動物が時間 C_i まで生存したという以上の情報は持たないと仮定することが、しばしば合理的である（p.243 第 2 パラグラフ）。しかし、もし打ち切りがその動物の運命と関係している場合、例えば発信器の不具合は研究対象が捕食された際に発信器が捕食者に破壊されたことによるとすれば、モデルに新たな情報を加えなければならない（p.243 第 3 パラグラフ）。

11.1.2 Interval Censoring, Staggered Entry, and Known Fates

今度は左打ち切りと中間打ち切りのデータのモデリングを扱う。まず、中間打ち切りのみの場合を考える。

（p.243 第 4 パラグラフは省略）

（p.243 第 5 パラグラフは省略）

全てが時間 t_0 から始まっている（左打ち切り無し）、寿命 T_i について独立な n 個のサンプルを持っているものとする。それぞれの個体がサンプリングの機会 $t_1, t_2, ..., t_k$ において生きているかどうかを決定することができるが、死んだ正確な時間はわからない。そのため、全ての観察は中間打ち切りである（p.243 第 6 パラグラフ）。

X_j を $j = 1, 2, ..., k, k+1$ について $[t_{j-1}, t_j)$ の間に死んだ個体数を示すものとする。ここで $t_{k+1} = \infty$ とすると、X_{k+1} は最後のサンプリング時点以降生残している個体の数となる。このデータは $X = (X_1, X_2, ...X_{k+1})'$ の多項ベクタ、およびベクタ $\boldsymbol{\pi}$ で要約される分割確率 π_j $(j = 1, 2, ..., k+1)$ で要約できる。モデルは $X \sim M_{k+1}(n, \pi)$ である（p.243 第 7 パラグラフ）。

（ここで多項分布の例を示す）

この式はとても一般的であり、寿命 T_i の分布に特別なモデルを仮定しなくてもパラメータ π_j について推論できるので、そういう意味でノンパラメトリックである。望むなら、π_i について、$f_\theta(\cdot)$ の分布系を特定して以下のようにモデリングすることで、寿命の分布に制約を与えることができる。

$$\pi_i = \int_{t_{i-1}}^{t_i} f_\theta(s) ds$$

この種類の制約は π_i の推論に必ずしも必要ではないが、推定を向上させると予想される（p.243 第 8 パラグラフ）。

続いて、左打ち切り（=時差加入; 個体群に研究の途中で個体が加入する）の場合を考える。全 n 個体が時間 t_0 から始まるという仮定に変わって、サンプリングの機会 $t_j, J = 0, 1, ..., k-1$ において n_j が時差加入すると仮定しよう。この後のモデルは容易に受け入れることができる。$X_{j,h}$ を、$h = j, j+1, ..., k$ に関して $[t_h, t_{h+1})$ の間の時間 t_j で死亡した個体の数を示すものとする。時間 t_j に時差加入した個体は少なくとも時間 t_j の終わりまでは生残しているので、$[t_h, t_{h+1})$ の期間に死亡する確率は $\pi_{j,h} = \pi_h/(1-\pi_j)$ である。$\boldsymbol{\pi_j} = (\pi_{j,j}, \pi_{j,j+1}, ..., \pi_{j,k})'$ とすると、時差加入のデータは $j = 0, 1, ..., k-1$ についての K 項の多項分布 $\boldsymbol{X_j} = (X_{j,j}, X_{j,j+1}, ..., X_{j,k})' \sim M_{k+1-j}(n_j, \boldsymbol{\pi_j})$

として記述できる（p.244 第 1 パラグラフ）。

しかし、研究によってはサンプリングの際にデータが得られない個体がある場合がある。このようなデータに対応するモデリングとして、Cormack-Jolly-Seber（CSJ）モデル（Cormack, 1964; Jolly, 1965; Seber, 1965）および Seber（1970）と Brownie *et al.*（1985）の標識再捕獲モデルにつながる（p.244 第 2 パラグラフ）。

11.2 OPEN POPULATION MARK-RECAPTURE – BAND-RECOVERY MODELS

ここから、標識再捕獲のモデル、そして標識再捕獲のデータに関するモデルを考えることで（標識再捕獲のモデルの）ベイズ的解析に目を向ける（p.244 第 3 パラグラフ）。

例として、7.3.1 で検討した研究、毎年 400 羽を（足輪をつけて）離した 3 年間の研究を再び検討する（p.244 第 4 パラグラフ）。

- 研究は 3 年間。毎年 400 羽の鳥に足輪をつけて放鳥
- その結果は死亡個体発見配列（Table 11.1）にまとめられている
- 各年はコホート i となる
- 太字は死亡個体数（後に出てくる統計量 r_{ij} に対応）
- コホートのサイズは R（この場合毎年 400）

問題となるのは、発見されなかった個体が死んでいるのか見つからないだけなのかわからないことである（Table 11.1 の斜体字。p.245 第 1 パラグラフ）

このデータをモデリングするためには、Table 11.2 にあるように全ての鳥の運命に関する情報を決定する必要がある。もしこの表を埋めることができれば、生残確率と再捕獲の確率に関する推論は前進するだろう（p.245 第 2 パラグラフ）。

- $\bar{r}_{ij}$：コホート i に属し期間 $(j, j+1]$ に死亡したが発見されなかった鳥の数
- w_j：生残している個体数
- $\sum_j \bar{r}_{ij} + w_j$：コホート i において発見されなかった個体数

A Reparameterization

再発見率を f_j、報告率を λ_j、生残率を S_j とする。その関係は以下のようである。

$$f_j = (1 - S_j)\lambda_j \tag{11.4}$$

ここで $0 < \lambda_j < 1$ である。報告率 λ_j は期間 $[j, j+1)$ に動物が死んだと報告される確率である。このパラメータ化の利点は、λ_j が生残確率とは機能的に独立であると言うことである（p.245 第 3 パラグラフ）。

Table 11.2 を $S'_j = 1 - S_j$ および $\lambda'_j = 1 - \lambda_j$ を用いて、生残確率と報告率で表現した物が Table 11.3 である（p.245 第 4 パラグラフ）。

Full Conditional Distribution for Latent Variable

Table 11.2 の統計量から、再捕獲されたかどうかにかかわらず期間 $[j, j+1)$ に死亡した標識個体の数を決定できる。この統計量を d_j とすると、

$$d_j = \sum_{i=1}^{j} (r_{ij} + \bar{r}_{ij})$$

これらの d_j 個の動物の内、再発見された動物の数は r_{ij} であり、d_j が与えられた下で r_{ij} は、試行数 d_j と生起確率 λ_j によって決定される二項確率変数である。この結果は、各コホートの $\{r_{ij}, \bar{r}_{ij}\}$ が多項分布であることと multinomial factorization theorem（Appendix B.7）によってもたらされる（p.246 第 1 パラグラフ）。

すべての動物が j 年の始めの時点で生存していると考えよう。この数を A_j とすると、それは以下のように与えられる。

$$\begin{aligned} A_j &= \sum_{i=1}^{j} (w_i + \sum_{h=j}^{k} (r_{ih} + \bar{r}_{ih})) \\ &= \sum_{i=1}^{j} (w_i + d_j) \end{aligned}$$

ここで k は調査の年数である。また、A_j は再帰的に以下のようにも定義できる（p.246 第 2 パラグラフ）。

$$A_j = \begin{cases} R_1 & j = 1 \\ A_{j-1} - d_{j-1} + R_j & j = 2, ..., k \end{cases}$$

例えば A_2 は Table 11.2 の統計量を足し合わせることで計算でき、その値は以下の表に示されている。また、太字は d_2 に寄与する、2 年目に死亡した鳥の個体数を示す。この表の統計量は、標識したことがわかっていて、2 年目のコホートの放鳥直後は生存している全ての鳥の数に対応する（p.246 第 3 パラグラフ）。

$\boldsymbol{r}_{12}$	r_{13}	$\bar{r}_{12}$	$\bar{r}_{13}$	w_1
$\boldsymbol{r}_{22}$	r_{23}	$\bar{r}_{22}$	$\bar{r}_{23}$	w_2

わずかな努力で、A_j、d_j はパラメータ $1 - S_j$ と指標 A_j で決定できる二項確率変数であることを示すことができる。

$\bar{r}_{ij}$ と w_j が既知の場合

S_j と λ_j の事後分布が直接得られる。例えば、S_j の事前分布が $\text{Be}(\alpha_S, \beta_S)$ であれば事後分布は $\text{Be}(\alpha_S + A_j - d_j, \beta_S + d_j)$ である。似たように、λ_j の事前分布が $\text{Be}(\alpha_\lambda, \beta_\lambda)$ であれば事後分布は $\text{Be}(\alpha_\lambda + \sum_{i=1}^{j} r_{ij}, \beta_\lambda + \sum_{i=1}^{j} \bar{r}_{ij})$ である。これらの確率密度は、もし完全なデータが得られればそれぞれの事後分布として正確に記述されるだろう（p.246 第 4 パラグラフ）。

$\bar{r}_{ij}$ と w_j が未知の場合

$\bar{r}_{ij}$ と w_j が観測されていない場合に、事後分布を得るためのギブズサンプラーの構築方法を示す。r_{11}、r_{12}、r_{13}、R_1、そしてパラメータが与えられた下で、未知パラメータのベクタ $(\bar{r}_{11}, \bar{r}_{12}, \bar{r}_{13}, w_1)'$ の分布は多項分布であり、それらの合計は放鳥後見つからなかったコホート 1 の鳥の総数に等しい。この多項分布の分割確率は、Table 11.3 における合計 1 になるように調整された各セルの確率に対応する。それは、$\bar{r}_{11}$ については $S_1'\lambda_1'/\psi_1$、$\bar{r}_{12}$ については $S_1S_2'\lambda_2'/\psi_1$、$\bar{r}_{13}$ については $S_1S_2S_3'\lambda_3'/\psi_1$、$w_1$ については $S_1S_2S_3/\psi_1$ であり、ここで $\psi_1 = S_1'\lambda_1' + S_1S_2'\lambda_2' + S_1S_2S_3'\lambda_3' + S_1S_2S_3$ である。類似した方法で、他のコホートに必要な多項分布を特定することができる（p.247 第 1 パラグラフ）

以下のようにして、ギブズサンプラーを構築する（p.247 第 2 パラグラフ）。

Step 1: それぞれのコホートにおいて発見されない個体数を $\bar{r}_{ij}$ と w_j に割り振ることで、完全な表（Table 11.2）を初期化する（どのような分配でもよい）

Step 2: $\mathrm{Be}(\alpha_S + A_j - d_j, \beta_S + d_j)$ 分布からサンプリングすることで S_j の完全な条件付分布からサンプルを得る

Step 3: $\mathrm{Be}(\alpha_\lambda + \sum_{i=1}^{j} r_{ij}, \beta_\lambda + \sum_{i=1}^{j} \bar{r}_{ij})$ 分布からサンプリングすることで、λ_j の完全な条件付分布からサンプルを得る

Step 4: 前の段階でサンプルされた S_j と λ_j の条件の下で適切な多項分布から全ての $\bar{r}_{ij}$ と w_j のサンプルを得る

Step 5: 2 から 4 を十分な回数繰り返す

試しに計算させてみた結果はこちら。

	Dies and not recovered			
i	1	2	3	Survives study
1	103	46	24	141
2		85	43	223
3			67	308

Constrained Models

このギブズサンプラーのすばらしい特徴は、$S_1 = S_2 = S_3 \equiv S.$（生残率が一定）あるいは $\lambda_1 = \lambda_2 = \lambda_3 \equiv \lambda.$（報告率が一定）といったような制限を容易に導入できる点である。例えば、ある報告率を当てはめたいときは、この制限の下では調査中に死亡し発見された動物の総数の完全な条件付分布が生起確率 $\lambda.$ を持った二項分布 $r.. = \sum_i \sum_j r_{ij}$ と調査期間に死亡した動物の総数 $\sum_j d_j$ で与えられるという事実を利用するだけでよい（p.247 第 3 パラグラフ）。

B.7 MULTINOMIAL DISTRIBUTION

二項分布は得られる結果が 2 つの事象を記述する; 得られる結果が 2 つ以上 $k \geq 2$ の事象は多項分布で記述される。これらの分布は生態学的な事象に適用する上で重要であり、足輪-放鳥や他の標識再捕獲のモデルにおいて中心的な役割を果たす。（これらの）研究において個々の動物は、捕獲と再発見に要約される遭遇の履歴と関連づけられている。研究の間に起こりうるそのような履歴は、有限回である。明確な履歴を持った動物の数を記述したベクタはしばしば、捕獲と再発見の過程を支配する無駄なパラメータと興味のある個体群動態のパラメータの関数の関数となっているパラメータのベクタを持つ多項確率変数としてモデリングされる。

$e_1, e_2, ..., e_k$ が確率変数 Υ から相互に排他的でかつ網羅的に得られるとし、$\Pr(\Upsilon = e_j) = \pi_j$ とする。π_j は非負で合計すると 1 になる。確率変数 Υ はカテゴリカルな分布を持つと言われる。Υ のサンプル $\Upsilon_1, \Upsilon_2, ..., \Upsilon_N$ はベクタ $\boldsymbol{X} = (X_1, X_2, ..., X_k)'$ で要約される。X_j は $\Upsilon_i = j$ となった個数である。

そのため、$\boldsymbol{X}$ は指標 N とパラメータのベクタ $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_k')$ をもつ k 項の多項分布といえ、$\boldsymbol{X} \sim M_k(N, \boldsymbol{\pi})$ と書ける。$\boldsymbol{X} = x = (x_1, x_2, ..., x_k)'$ が得られる確率は、以下のようである。

$$\Pr(\boldsymbol{X} = x) = \frac{N!}{\prod_{j=1}^{k} x_j!} \prod_{j=1}^{k} \pi_j^{x_j}$$

$\boldsymbol{X}$ の平均値は驚く事なかれ、$N\boldsymbol{\pi}$ なのである。$\boldsymbol{X}$ の分散共分散行列は、$i \neq j$ の場合は i 行 j 列においては $-N\pi_i\pi_j$ であり、対角線上の要素は $N\pi_i(1 - \pi_i)$ である。

多項分布のモデリングは、共変量 $\boldsymbol{z}$ を観測した下での$\boldsymbol{\pi}$ に関するパラメトリックなモデルの特定を含んでいる。

多項確率変数はポアソン確率変数と重要な関係がある。ベクタ $\boldsymbol{A} = (A_1, A_2, ..., A_n)'$ が独立したポアソン確率変数から得られ、i 番目の要素の平均は μ_i とする。これらの合計が $T = \sum_i A_i$ とすると、ベクタ $\boldsymbol{A}$ は多項分布 $\boldsymbol{A}|T \sim M_n(t, \boldsymbol{\pi})$ であり、

$$\pi_i = \frac{\mu_i}{\sum_j \mu_j} \tag{B.11}$$

特に、$\mu_i = \lambda q(\boldsymbol{\theta}; z_i)$ とする; ここで、λ は基礎となる率であり、$q(\boldsymbol{\theta}; z)$ はその率に共変量が与える影響を記述した関数である。動物の個体数指数カウントはしばしばこのような方法で記述され、λ は（動物の）基礎的な量を示し $q(\boldsymbol{\theta}; z)$ は発見率を示し、それぞれ動物個体群の動態と観察者の効果を示す。全体の調整（Conditioning）は、モデルからパラメータ λ を除く効果がある。それは、式 B.11 の右辺の分子と分母を相殺する。

Mutinomial Factorization Theorem

$e_1, ..., e_k$ について、興味が部分集合 $S = \{e_j\}_{j \in K}$ に限定されているとする。ここで $K \subseteq \{1, 2, ..., k\}$ である。$\boldsymbol{X}_S$ は $\boldsymbol{X}$ の対応する部分の部分集合を示すとする。全体が与えられた下で

$\boldsymbol{X}_S$ の要素の同時確率もまた多項分布である。

$$\Pr(\boldsymbol{X}_S = x_S | N_S) = \frac{N_S!}{\prod_{j \in K} x_j!} \prod_{j \in K} \gamma_j^{x_j}$$

ここで

$$N_S = \sum_{j \in K} X_j$$

また

$$\gamma_j = \frac{\pi_j}{\sum_{h \in K} \pi_h}, \quad j \in K$$

Example

$\{1, 2, 3, 4, 5\}$ のいずれかの値を取り、その確率が $\pi_1, ..., \pi_5$ であるカテゴリー確率変数の観測値を、N 回試行して得たとする。もし興味が 4 未満に限定されていれば、$\boldsymbol{X}_S(X_1, X_2, X_3)'$ の同時確率 $X_1 + X_2 + X_3 = T$ は、

$$\Pr(\boldsymbol{X}_S = (x_1, x_2, x_3)' | T) = \frac{T!}{x_1! x_2! x_3!} \prod_{j=1}^{3} \left(\frac{\pi_j}{\pi_1 + \pi_2 + \pi_3} \right)^{x_j}$$

B.8 EXPONENTIAL DISTRIBUTION

X が以下の密度関数を持ち、$X \sim E(\lambda)$ でパラメータ $\lambda > 0$ を持つ場合、指数関数という。

$$f(t) = \lambda \exp(-\lambda t), t > 0 \tag{B.12}$$

平均と分散はそれぞれ $1/\lambda$ および $1/\lambda^2$ である。指数確率変数 X のモードは 0、中央値は $\ln(2)/\lambda$ である。

指数分布は、ポアソン過程に従う滅多に起こらない事象の間の待ち時間としてしばしば記述される。$\{N(t); t > 0\}$ が頻度 λ のポアソン過程に従う、つまり $N(t) \sim P(\lambda t)$ とする。X を最初の事象が起こるまでの時間とし、$F(t)$ が X の累積密度関数とすると、

$$Pr(X > t) = 1 - F(t) = Pr(N(t) = 0) = \frac{(\lambda t)^0 \exp(-\lambda t)}{0!} = \exp(-\lambda t)$$

そのため、

$$F(t) = 1 - \exp(-\lambda t) \tag{B.13}$$

B.13 の両辺を t で微分すると B.12 が得られる。

指数確率変数は一様乱数が与えられた下で容易に生成できる。$U \sim U(0, 1)$ であれば、$X = -\log(U)/\lambda \sim E(\lambda)$ である。

たくさんの独立な指数確率変数の最小値は指数確率変数である。$X_i \sim E(\lambda_i), i = 1, 2, ..., n$ は独立であるとする。すると、

$$Y = \min\{X_1, X_2, ..., X_n\} \sim E(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$$

さらに、$Pr(X_i = \min\{X_1, X_2, ..., X_n\}) = \lambda_i / \sum_j \lambda_j$ である。

Lack of Aging Property

指数分布は、数々の他の分布にはない注目すべき特徴を持っているので、生残分析において主要な役割を果たしているが、それらの特徴の中でも主要なものは加齢の影響が無いことである。t と h を非負の数とし、X を以下のような特性を持つ確率変数とする。

$$\Pr(X > t + h | X > t) = \Pr(X > h) \tag{B.14}$$

時間 t まで生残した上でさらに時間 h 生残する確率は、ある初期時点から時間 h だけ生残する確率に等しい。もし人間の寿命が指数分布に従うのであれば、50 歳の人間が 55 歳まで生残する割合、10 歳の人間が 15 歳まで生残する割合、新生児が 5 歳まで生残する割合は全て等しい。

$X \sim E(\lambda)$ であれば式 B.14 が正しいことを証明するのは簡単である。生残関数 $S(t) = 1 - F(t)$ を用いると、式 B.14 は以下のように書ける。

$$\frac{S(t + h)}{S(t)} = S(h) \tag{B.15}$$

式 B.15 の両辺に $S(t)$ をかけ、その後両辺から $S(t)$ を引き、h で割ると、以下の式が得られる。

$$\frac{S(t + h) - S(t)}{h} = S(t) \left(\frac{S(h) - 1}{h} \right) \tag{B.16}$$

$S(0) = 1$ なので、式 B.16 の右辺の 1 を $S(0)$ で置き換えることができる。そのため、式 B.16 において h について 0 の極限を両辺で取ると、以下の式が得られる。

$$S'(t) = S(t)S'(0) \tag{B.17}$$

$S'(0)$ は定数である; これを λ と呼ぶ。そのため、式 B.17 は単純な微分方程式と見なせ、その解は以下のようである。

$$S(t) = \exp(-\lambda t)$$

よって、式 B.14 の加齢の影響がない特性は指数分布にのみ対応する。

Hazard Functions

加齢の影響がないという性質は、ハザード関数や「死に追いやる力」関数においてしばしば記述されている。ある齢 t まで生残した個体が、次のある期間 $(t, t + h]$ に死亡する確率を考える。生残は連続的な確率変数とすると、期間 $(t, t + h]$ に死亡する確率は $h \to 0$ となると 0 に近づく; し

かし、瞬時的な（死亡）率を算出するために、期間の長さで確率を規格化するかもしれない。この死亡率は t の関数として変動する。そのため、ハザード関数は以下のように定義される。

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0}\frac{1}{h}(Pr)(X \in (t, t+h]|X > t) \\ &= \lim_{h\to 0}\frac{F(t+h)-F(t)}{h\ S(t)} = \frac{f(t)}{S(t)}\end{aligned}$$

指数分布ではハザード関数は一定、すなわち $\lambda(t) \equiv \lambda$ である; 加齢の影響がないという性質は「死に追いやる力」が一定であることと等価である。

ハザード関数は分布関数から計算でき、逆も可能である。そのためハザード関数は $\lambda(t) = f(t)/S(t)$ と定義され、

$$\int_0^x \lambda(t)dt = \int_0^x \frac{f(t)dt}{S(t)} = -\log(S(x))$$

よって、

$$S(x) = \exp(-x\Lambda(x))$$

ここで、

$$\Lambda(x) = \frac{1}{x}\int_0^x \lambda(t)dt$$

は期間平均のハザード関数である。この、ハザード関数と分布関数の 1 対 1 の対応は、生残分布の特徴がハザード関数で表現可能であり、その逆も成り立つことを示しているかも知れない。そのため、ハザード関数一定である指数分布は生残パターンを調査する、もっとも単純な基礎である。例えば、人間の生残は「バスタブの形をした」ハザード関数でしばしば表現される（Fig. B.4）:「死に追いやる力」は新生児と老齢で最も高くなり、幼児期に入ると減少し、老齢期に入ると増加する。

Weibull Distribution

もし $X \sim E(\lambda)$ で $\alpha > 0$ が定数であれば、$\Upsilon = X^{1/\alpha}$ はワイブル分布 $\Upsilon \sim W(\lambda, \alpha)$ を持つ。ワイブル分布のハザード関数は $\lambda_w(t) = (\alpha\lambda)t^{(\alpha-1)}$ であり、$\alpha < 1$ であれば必ず減少し、$\alpha > 1$ であれば必ず増加する。