ハザード関数の統計解析と生命表

松下敬一郎・稲葉　寿[1)]

日常の生活において，非可逆的な事象は数多くみられる．ヒトの死やマウスの投薬実験はその典型的な例であるが，カメラを繰り返し落としてレンズを割ってみたり，電球のフィラメントが切れたり，金魚すくいの紙が破れるといった例があげられる．このように時間や試行回数に伴って発現する非可逆的な事象の背後に存在すると思われる確率法則を研究するために，ハザード関数を用いた統計解析が行われている[2)]．ハザード関数の定義については後述するが，生命表の諸変数はハザード関数（生命表の死力に対応する）によって書き表わされることができる．従って，ハザード関数の統計解析の手法の発展に呼応して，「死亡の秩序」に関する分析はより高度な手法により展開されるようになり，その応用範囲も拡張されたのである．とりわけデータの観察が部分的に中断された場合の分析には有用である[3)]．本論では，まずハザード関数の基礎理論を説明し生命表の諸変数とそれとの関係を明示する．次に人口学の分野でそれを応用するための基本理論を紹介する．

I　ハザード関数の基礎理論と生命表の諸変数

1. 確率密度関数

まず，｛x才で死亡する｝という事象をXで表すことにする．Xは連続分布に従い，その定義域として0以上の実数を仮定する．さらに，事象X[4)]が時間xと$x+\Delta x$のあいだに生じる確率，つまりx才から$x+\Delta x$才のあいだに死亡が発生する確率を，Pr（$x \leqq X < x+\Delta x$）と表す．そこで，Xに関する確率密度関数fは式(1)のように表せる[5)]．

1）本論は経済人口学および数理人口学の各分野を専門とする筆者等が互いに関連する領域であるハザード関数を用いる統計解析についてその方法論をサーベイする過程でまとめたものである．本稿を作成するにあたり小林和正，南条善治，五十嵐忠孝，河邊宏，廣嶋清志，伊藤達也の各氏から貴重なコメントをいただいた．ここに記して感謝したい．

2）修復可能な場合，つまり個体Iが状態A（健康）から状態B（病気）に推移する事象を分析する際に状態Bから状態Aに戻る（健康を回復した）個体を含めた場合，の分析も可能である．さらに推移前および推移後の状態が唯一ではない場合，例えば，個体Iが状態A（無配偶）から生涯を始めるものの明確に区別できる複数の状態（状態B（有配偶），状態C（離別），状態D（死別））の間を推移する場合，についても分析が可能である．F. J. Willekens, I. Shah, J. M. Shah, and P. Ramachandran, "Multi－state Analysis of Marital Status Life Tables：Theory and Applications", *Population Studies*, Vol. 36, No. 1, 1982, pp. 129－144.

3）J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, New York, John Wiley and Sons 1980, および J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, New York, John Wiley and Sons 1982, 等に数値計算例やより詳しい理論的展開が示されている．

4）厳密には，事象Xではなく確率変数Xあるいは確率事象Xとするべきであるが，誤解が生じると思われないかぎり略している．

5）従って，事象Xが時間xと$x+\Delta x$のあいだに生じる確率はfを用いて

$$\Pr\ (x \leqq X < x+\Delta x) = \int_{x}^{x+\Delta x} f\,(w)\,dw$$

と表される．

$$f(x) = \lim_{\Delta x \to 0} \frac{\Pr\ (x \leqq X < x + \Delta x)}{\Delta x} \quad \cdots\cdots (1)$$

これはx才の瞬時に発生する死亡の確率密度を示しており，後述するとおり生命表関数の単位時間当たり死亡数の極限値，$\lim_{n \to 0} n^{d}x / n$に対応する．便宜上，生命表関数の数値は出生時人口（l_0）を100,000人あるいは適当な数値に置き換えて表されているが，ハザード関数と対応させて考える場合，l_0 を一定乗数あるいは1と考えればよい[6]．

2. 生残確率関数

x才まで生き残る確率Fは式(2)で示すように事象Xがx時点以後に生じる（x才以上で死亡する）確率Fと同じであり，生命表関数の l_x に対応する．

$$F(x) = \Pr\ (x \leqq X) \quad \cdots\cdots (2)$$

ここで，Fは連続関数で微分可能であると仮定する．確率密度関数fと生残確率関数Fとの関係は

$$f(x) = -\frac{d\,F(x)}{d\,x}$$

で示される[7]．上式をFについて解くと，式(3)に示すように，Fは1とXの累積確率との差に等しくなる．

$$F(x) = \int_{x}^{\infty} f(w)\,dw = 1 - \int_{0}^{x} f(w)\,dw \quad \cdots\cdots (3)$$

ここで，確率分布の定義上，$F(0) = 1$ と $F(\infty) = 0$ が満たされている．つまり，出生時の生残確率は1であり，経過時間が十分に長ければその出生コーホートの人口はすべて死亡し，生残確率は0となる．

3. ハザード関数

ここで，時間xまで生残するという条件のもとで事象Xが時間xと$x + \Delta x$のあいだに生じる条件

6）本論では生命表をある個体がもつ確率法則と考えてハザード関数と対応させている。ある人口が経験する確率過程としての生命表とハザード関数とをより厳密に対応させるため，コーホート生命表の生残数 l_x について考えよう．均質な人口の生残確率があるハザード関数で定まると仮定すると，実際の生残数 l_x は成功の確率が生残確率$F(x)$となる二項目分布に従う．まず，ここで l_x と$F(x)$との対応関係がある．
さらに，l_x の確率分布は

$$\Pr\ (l_x = k) = \frac{l_0\ !}{k!\,(l_0 - k)!}\,F(x)^{k}\,(1 - F(x))^{l_0 - k}$$

となり，l_x の期待値は$E\,[\,l_x\,] = l_0 \cdot F(x)$ となることから，生残数の期待値は，l_0 と生残確率との積に等しい．ここで，l_x は期待値関数Eをつうじて$F(x)$と対応している．

7）x才から$x + \Delta x$才のあいだに事象Xが生じる確率は，x才以上でXが生じる確率と$x + \Delta x$才以上でXが生じる確率の差として表すことができるから，$F(x)$と$F\ (x + \Delta x)$の差に等しく，

$$\Pr\ (x \leqq X < x + \Delta x) = \Pr\ (x \leqq X) - \Pr\ (x + \Delta x \leqq X)$$
$$= F(x) - F(x + \Delta x)$$

と表せる。これを式(1)に代入すれば求まる．

付確率について考えよう．これはx才まで生きた人間が次のΔx年の間に死亡する確率である．この条件付確率は$\Pr(x \leqq X < x+\Delta x \mid x \leqq X)$と表される．$X$に関するハザード関数$\lambda$は式(4)で定義される．

$$\lambda(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \leqq X < x+\Delta x \mid x \leqq X)}{\Delta x} \quad \cdots\cdots (4)$$

これはx才の瞬時に発生する死亡の条件付確率の密度を示しており，生命表関数の死力μ_xに対応する．式(4)は次式のように確率密度関数fと生残率関数Fの比に書き直すことができる[8]．

$$\lambda(x) = \frac{f(x)}{F(x)} = -\frac{d}{dx}(\ln F(x))$$

両辺を0からxまで積分すると，生残確率関数Fと確率密度関数fはλの関数として式(5)および式(6)のように表すことができる[9]．

$$F(x) = \exp\left[-\int_0^x \lambda(w)\,dw\right] \quad \cdots\cdots (5)$$

$$f(x) = \lambda(x) \cdot \exp\left[-\int_0^x \lambda(w)\,dw\right] \quad \cdots\cdots (6)$$

従って，ハザード関数$\lambda(x)$を用いてXの分布を示すことができるのである．λはfやFと同等の役割を果たし，λの理論分布およびλに関する実験データや調査データからXの分布についての情報が得られる．

4. その他の関数

(1) 平均余命（条件付期待値）

生命表関数の平均余命$\mathring{e}_x$は，x才に達した人（x才以上で死亡する場合）の残された人生の年月の条件付期待値を示すものである．それは，事象Xがx時点以後に生じる場合の$X-x$の条件付期待値であるから，$r(x) = \mathrm{E}[X-x \mid x \leqq X]$　と表すことができる．部分積分を求めてrをλの関数で表すと

8）ここで，{x才から$x+\Delta x$才のあいだに死亡する}という事象をAとし，{x才以上で死亡する}という事象をBとすると，ここで扱っている条件付確率は，

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

で表される。ところが$A \subset B$であるから$A \cap B$はAに等しくなり，

$$\Pr(A \mid B) = \frac{\Pr(A)}{\Pr(B)}$$

となる。従って，分子は

$$\Pr(x \leqq X < x+\Delta x \mid x \leqq X) = \frac{\Pr(x \leqq X < x+\Delta x)}{\Pr(x \leqq X)}$$

と書くことができる．これを式(4)に代入すれば求まる．

9）$\exp[\cdot]$は指数関数を表す。

式(7)が求まる[10].

$$r(x) = \frac{1}{\exp\left[-\int_0^x \lambda(w)\,dw\right]} \cdot \int_x^\infty \exp\left[-\int_0^w \lambda(u)\,du\right] dw \quad \cdots\cdots (7)$$

なお，$\mathring{e}_x = T_x / l_x$ であり，生存延べ年数 T_x に対応する部分 $\bar{T}_x$ は式(8)で表される．

$$\bar{T}_x = \int_x^\infty \exp\left[-\int_0^w \lambda(u)\,du\right] dw \quad \cdots\cdots (8)$$

(2) 期間変数：${}_nd_x$，${}_nq_x$，${}_nL_x$ について

生命表における死亡数 ${}_nd_x$，死亡確率 ${}_nq_x$，静止人口 ${}_nL_x$ に対応する変数は，それぞれ式(9)，(10)，(11)で示されるように λ の関数として表すことができる．死亡数 ${}_nd_x$ は x 才から $x+n$ 才のあいだに生じる生命表上の死亡数，あるいは出生時人口に等しい死亡数全体に占める該当年齢の死亡の割合である．これは，注4）に示されるように，時間 x から $x+n$ のあいだに事象 X が生じる確率，つまり密度関数を x から $x+n$ まで積分したものと対応する．

$${}_n\bar{d}_x = \int_x^{x+n} \lambda(w) \cdot \exp\left[-\int_0^w \lambda(u)\,du\right] dw \quad \cdots\cdots (9)$$

生命表上の死亡確率 ${}_nq_x$ は x 才から $x+n$ 才のあいだに死亡する人口の x 才時人口に対する比である．これは事象 $\{x \leqq X\}$ が生じる場合に $\{x \leqq X < x+n\}$ が生じる条件付確率 $\Pr(x \leqq X < x+n \mid x \leqq X)$ と対応する．

$${}_n\bar{q}_x = \frac{\int_x^{x+n} \lambda(w) \cdot \exp\left[-\int_0^w \lambda(u)\,du\right] dw}{\exp\left[-\int_0^x \lambda(w)\,dw\right]} \quad \cdots\cdots (10)$$

---

10) $$r(x) = \int_x^\infty (w-x)\frac{f(w)}{F(x)}\,dw$$

$$= \frac{1}{F(x)} \cdot \left\{\left[-(w-x)F(w)\right]_x^\infty + \int_x^\infty F(w)\,dw\right\}$$

$$= \frac{1}{F(x)} \cdot \int_x^\infty F(w)\,dw$$

$$= \frac{1}{\exp\left[-\int_0^x \lambda(w)dw\right]} \cdot \int_x^\infty \exp\left[-\int_0^w \lambda(u)\,du\right] dw$$

となり，式(7)が導かれる．

表1　変数対照表

生命表関数表示		確率統計理論表示		λによる表示
死力	μ_x, $\lim_{n\to 0}\frac{{}_nq_x}{n}$	$\lambda(x)$	$\lim_{\Delta x\to 0}\frac{\Pr(x\leqq X<x+\Delta x \mid x\leqq X)}{\Delta x}$	$\lambda(x)$
生存率	l_x	$F(x)$	$\Pr(x\leqq X)$	$\exp\left[-\int_0^x\lambda(w)\,dw\right]$
（密度関数）	$\lim_{n\to 0}\frac{{}_nd_x}{n}$	$f(x)$	$\lim_{\Delta x\to 0}\frac{\Pr(x\leqq X<x+\Delta x)}{\Delta x}$	$\lambda(x)\exp\left[-\int_0^x\lambda(w)\,dw\right]$
平均余命	$\mathring{e}_x$	$r(x)$	$\mathrm{E}\left[X-x \mid x\leqq X\right]$	$\dfrac{\int_x^\infty \exp\left[-\int_0^w\lambda(u)\,du\right]dw}{\exp\left[-\int_0^x\lambda(w)\,dw\right]}$
生存延べ年数	T_x	$\int_x^\infty F(w)\,dw$	$\mathrm{E}\left[X-x \mid x\leqq X\right]\cdot\Pr(x\leqq X)$	$\int_x^\infty \exp\left[-\int_0^w\lambda(u)\,du\right]dw$
死亡数	${}_nd_x$	$\int_x^{x+n} f(w)\,dw$	$\Pr(x\leqq X<x+n)$	$\int_x^{x+n}\lambda(w)\cdot\exp\left[-\int_0^w\lambda(u)\,du\right]dw$
死亡確率	${}_nq_x$, $\frac{{}_nd_x}{l_x}$	$\dfrac{\int_x^{x+n} f(w)\,dw}{F(x)}$	$\Pr(x\leqq X<x+n \mid x\leqq X)$	$\dfrac{\int_x^{x+n}\lambda(w)\cdot\exp\left[-\int_0^w\lambda(u)\,du\right]dw}{\exp\left[-\int_0^x\lambda(w)\,dw\right]}$
静止人口	${}_nL_x$	$\int_x^{x+n} F(w)\,dw$	$\mathrm{E}\left[X-x \mid x\leqq X<x+n\right]\cdot\Pr(x\leqq X<x+n)$ $+n\cdot\Pr(x+n\leqq X)$	$\int_x^{x+n}\exp\left[-\int_0^w\lambda(u)\,du\right]dw$

静止人口，あるいは x 才から $x+n$ 才のあいだの生存延べ人口 ${}_nL_x$ は生残確率を x から $x+n$ まで積分したものに対応する．

$$ {}_n\bar{L}_x = \int_x^{x+n} \exp\left[-\int_0^w \lambda(u)\,du\right] dw \quad \cdots\cdots (11) $$

なお，${}_n\bar{q}_x$ と ${}_n\bar{d}_x$ の期間 n を 0 に近づけた場合における単位時間あたりの極限値はそれぞれ $\lambda(x)$ と $f(x)$ となる．

$$ \lim_{n\to 0} \frac{{}_n\bar{q}_x}{n} = \lambda(x) $$

$$ \lim_{n\to 0} \frac{{}_n\bar{d}_x}{n} = f(x) = \lambda(x) \cdot F(x) $$

5. 離散分布の場合

離散分布の場合についても連続分布の場合と同様に生命表関数をハザード関数に対応させることができる．変数値が小さいものから順番に $x_1 < x_2 < \cdots$ と並べてあるものとする．ここで，分布確率関数と生残確率関数をそれぞれ式(12)，式(13)とする．

$$ f(x_i) = \Pr(X = x_i) \quad \cdots\cdots (12) $$

$$ F(x_t) = \Pr(x_t \leqq X) = \sum_{j \mid x_t \leqq x_j} f(x_j)^{11)} \quad \cdots\cdots (13) $$

この場合，ハザード関数は，事象 X が時点 x_j 以後に生じる場合に事象 X がちょうど時点 x_j で生じる確率であるから，式(14)で表される．

$$ \lambda_j = \Pr(X = x_j \mid x_j \leqq X) = \frac{f(x_j)}{F(x_j)} \quad (j = 1, 2, \cdots\cdots) \quad \cdots\cdots (14) $$

ハザード関数を用いて生残関数および分布関数を表すと，式(15)，式(16)のようになる[12)]．

$$ F(x_t) = \prod_{j \mid x_t > x_j} (1 - \lambda_j)^{13)} \quad \cdots\cdots (15) $$

---

11) 右辺は j 番目の x の値が x_t の値以上のものについて $f(x_j)$ をたしあわせることを意味している．

12) F の定義より，$F(x_{t+1}) = F(x_t) - f(x_t)$ が成り立つ．さらに，

$$ 1 - \lambda_j = \frac{F(x_j) - f(x_j)}{F(x_j)} = \frac{F(x_{j+1})}{F(x_j)} $$

が成り立つ。ゆえに，

$$ F(x_t) = \prod_{j=1}^{t-1} \frac{F(x_{j+1})}{F(x_j)} = \prod_{j \mid x_t > x_j} (1 - \lambda_j) $$

が示される．

13) 右辺は j 番目の x の値が x_t の値未満のものについて $1 - \lambda_j$ をかけあわせることを意味している．

$$f(x_j) = \lambda_j \cdot \prod_{i=1}^{j-1} (1-\lambda_i) \quad \cdots\cdots (16)$$

実験や調査データから簡単に生残確率$F(t)$の推定値を直接求めるProduct Limit（あるいはKaplan－Meier）推定値は最尤法推定値と等しくなることが知られてる[14]．Product Limit推定値$\hat{F}(t)$は式(17)で表される[15]．

$$\hat{F}(t) = \prod_{j \mid t > t_j} \left(1 - \frac{d_j}{n_j}\right) \quad \cdots\cdots (17)$$

ここでハザード率λ_jは

$$\lambda_j = \frac{\Pr(T = t_j)}{\Pr(t_j \leqq T)}$$

であるから，分母にはj期まで生き残っているサンプルn_jが対応し，分子にはj期で死亡したサンプルd_jが対応している．この場合，観察対象の脱落を考慮する上で，d_jは一時点の死亡数を意味するものであり，ある期間の死亡数ではないことに注意する必要がある．

以上では，第一にハザード関数が生命表の死力と同一であることを述べた．第二に生命表の諸関数の確率表示を試みた．第三にハザード関数（死力）を用いて生命表の諸関数を導いた．これらの結果は対照表として表1に再掲してある．このように生命表と不可分の関係にあるハザード関数は，人口に関するデータの統計分析の際にどのように利用できるのであろうか．次に，指数分布，正規分布，比例ハザードモデル，Multiple Decrement 生命表[16]へ応用するための基本理論を紹介する．

Ⅱ　ハザード関数の応用理論

1. 指数分布

ここで，ハザード率が一定の場合の統計分析について考えよう．二つの仮設的な例をあげることができる．まず，つがいの形成が全く無作為で人口密度のみによって相手に出会う確率が定まるものと仮定する．実験的な状況下で性比1の無配偶個体群の人口密度が一定に保たれているとするならば，つがい形成の確率は一定となる．ある時点でこの空間に放出された無配偶個体群がつがいを形成する過程は一定ハザード率の分布に従う．

14）Lawless 前掲書（注3）74－76ページ参照．

15）$F(t)$の分散の漸近推定値として次式が一般に用いられている．

$$\widehat{\mathrm{Var}}(\hat{F}(t)) = (\hat{F}(t))^2 \sum_{j \mid t > t_j} \frac{d_j}{n_j (n_j - d_j)}$$

従って，95％の信頼区間は

$$\hat{F}(t) \pm 1.96 \left[\widehat{\mathrm{Var}}(\hat{F}(t))\right]^{1/2}$$

で近似される．

16）竹中規雄はMultiple Decrement Table を多重脱退残存表と訳している．新生命保険実務講座刊行会編，『新生命保険実務講座第7巻数理』46ページ参照．意訳としては多死因生命表あるいは複合生命表が考えられる．

次に，ジョブサーチ理論を応用したKeeley（1977）[17]の結婚モデルを考えよう．結婚に伴う夫婦家計の総生産から配偶者候補の取り分を差し引いた部分がジョブサーチモデルの提示賃金にあたる．この賃金にあたる部分の分布が一定で，必ず提示賃金が得られるものとすると，一回のサーチにおいてそれがある一定の範囲内に入る確率も一定となる．このような結婚市場に参入した独身者が結婚していく過程も一定ハザード率の分布に従う．

このような場合，時間tまで独身で時間tで結婚するハザード率は時間にかかわらず一定で$\lambda(t)=\lambda$となる．連続分布を仮定して式(5)にλを代入して生残確率を求めると，$F(t)=\exp[-\lambda t]$となる．従って，確率密度関数および期待サーチ期間はそれぞれ$f(t)=\lambda\cdot\exp[-\lambda t]$，$r(t)=1/\lambda$となる[18]．これは$T$が指数分布に従うことを示している．なお，指数関数モデルが現実に適合するかどうかを経験的に確かめるためには，生残確率の推定値の対数値$\ln\hat{F}(t)$と時間tとをプロットすればよい．もしも指数分布に従うならば，原点を通る直線で近似される．

後述の比例ハザードモデルとの類似性を考える上で，ここで，指数分布に対数変換をおこなってみよう．一般に分布Tを分布Yに変換する場合（$T\xrightarrow{g}Y$），両者の確率密度分布の関係について，式(18)が成り立つ[19]．

$$f_Y(y)=f_T(g^{-1}(y))\cdot\frac{d}{dy}(g^{-1}(y)) \quad \cdots\cdots (18)$$

そこで，$Y=g(T)=\ln T+\ln\lambda$　の変数変換をおこなうと，Yの理論分布は式(19)で表されるように最小値の分布になる[20]．

$$f(y)=\exp\left[y-e^{y}\right],\quad y\in(-\infty,\infty) \quad \cdots\cdots (19)$$

次に，実験あるいは調査データからλを推定する方法を考えよう，まず，サンプル数がnで結婚までの待ち時間tをすべて記録したデータがあるとしよう．この場合，尤度関数は式(20)となる．

$$L(\lambda)=\prod_{i=1}^{n}f(\lambda;t_i)=\lambda^n\cdot\exp\left[-\lambda\sum_{i=1}^{n}t_i\right] \quad \cdots\cdots (20)$$

従って，Lを最大にするλの値$\hat{\lambda}$は式(21)で得られ，

$$\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}t_i} \quad \cdots\cdots (21)$$

---

17) M. C. Keeley, "The Economics of Family Formation", *Economic Inquiry*, Vol. 15, No. 2, 1977, pp. 238–250.

18) この場合，結婚市場に参入してから結婚するまでの待ち時間の平均はr年となる．

19) P. J. Bickel and K. A. Doksum, *Mathematical Statistics*, Holden - Day, San Francisco 1977, pp. 9–12.

20) この変数変換の場合，

$$g^{-1}(y)=\frac{d}{dy}(g^{-1}(y))=\exp\left[y-\ln\lambda\right]$$

を代入して$f_Y(y)$が求まる．

待ち時間の期待値は $1/\hat{\lambda}$ となる[21]. この結果，全サンプルが観測された場合には，待ち時間の期待値と単純平均とは等しくなる.

一方，nサンプルのうちd番目の結婚までの待ち時間(t_d)が観測されたデータがあるとしよう. この場合$n-d$個のサンプルはセンサーされており[22]，待ち時間がt_d以上であることしかわかっていないため，このサンプルが尤度に与える貢献部分は$F(t_d)$に等しく$\exp[-\lambda t_d]$となる. さらに，待ち時間を観測したd個のサンプルの尤度の貢献は各々の待ち時間t_iについて$f(t_i)$に等しく$\lambda\exp[-\lambda t_i]$となる. 従って，尤度は式(22)となる[23].

$$L(\lambda)=\prod_{i=1}^{n} f(t_i)^{\delta_i}\cdot F(t_d)^{1-\delta_i}$$
$$=\lambda^d\cdot\exp\left[-\lambda\left(\sum_{i=1}^{d} t_i+(n-d)t_d\right)\right] \quad \cdots\cdots (22)$$

ただし，δ_iはセンサーの有無を示す指標でデータが観測された場合は$\delta_i=1$であり，センサーされた場合は$\delta_i=0$である. ゆえにLを最大にするλの値$\tilde{\lambda}$は式(23)で得られ，

$$\tilde{\lambda}=\frac{d}{\sum_{i=1}^{d} t_i+(n-d)t_d} \quad \cdots\cdots (23)$$

この不完全なサンプル観測の場合の待ち時間の期待値は$1/\tilde{\lambda}$となる[24].

2. 正規分布

初潮年齢が二つのパラメーターμとσ^2で決まる正規分布に従うことが認められている[25]. ここでは，部分的にセンサーされたサンプルデータからμとσ^2を推定する方法を考えよう. 標準化した分布密度関数$\phi(z)$および生残確率関数$Q(z)$をそれぞれ

---

21) $T=\sum_{i=1}^{n} t_i$とするとT/rはパラメータnのガンマ分布を持つことが知られており，$2T/r$は自由度$2n$のカイ二乗分布χ^2_{2n}に従う. この性質を用いてrについての仮説検定や信頼区間の計算を行うことが出来る. この場合，rの信頼区間はαの有意水準について

$$\frac{2T}{\chi^2_{2n,1-\alpha/2}}\leqq r\leqq\frac{2T}{\chi^2_{2n,\alpha/2}}$$

となる. さらに，二種のサンプルデータから得られたrを比較するためには尤度比検定が利用できる.

22) この場合，" censor " とはサンプルの脱落や実験の打ち切りを意味するが，適訳がないため，本論ではそのまま使用している. 例えば，一定期間についてのみ調査が行われた場合，調査以前および調査以後のデータを入手することはできず，センサーされたという.

23) 異なった方法でセンサーされたサンプルを含む場合の尤度関数の導き方については Lawless 前掲書(注3) 31-41ページを参照されたい.

24) この場合には

$2(\sum_{i=1}^{d} t_i+(n-d)t_d)/r$がカイ二乗分布$\chi^2_{2d}$に従う.

25) D. C. Wilson and I. Sutherland, " Further Observations on the Age of the Menarche," British Medical Journal, Vol. 2, 1950, pp. 862-866. 箕輪真一，平木陽一，滝川弘志，「最近の女子思春期の発育に関する研究」，『民族衛生』37巻3号，1971年，pp. 113-125参照.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \quad exp\left[\frac{-z^2}{2}\right]$$

$$Q(z) = \int_z^{\infty} \phi(x)\, dx$$

で示す．一般に，平均μと分散σ^2の正規分布を持つ確率事象$T \sim N(\mu, \sigma^2)$の確率密度関数と生残確率関数をそれぞれ式(24)，式(25)で表すことができる．

$$f(t) = \frac{1}{\sigma}\ \phi\ \left(\frac{t-\mu}{\sigma}\right) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (24)$$

$$F(t) = Q\ \left(\frac{t-u}{\sigma}\right) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (25)$$

初潮が観測されたデータを集合D，そのサンプル数をdとし，センサーされたデータを集合をCとする．t_iを初潮年齢あるいはセンサー時の年齢とする．尤度関数は式(26)で与えられる．

$$L(\mu, \sigma\ ;\ t_i) = \prod_{i \in D} \frac{1}{\sigma}\ \phi\ \left(\frac{t_i-\mu}{\sigma}\right) \cdot \prod_{i \in C} Q\ \left(\frac{t_i-\mu}{\sigma}\right) \quad \cdots\cdots\cdots\cdots \quad (26)$$

両辺の対数をとると，対数尤度関数は定数項を省略して

$$\ln L = -\ d \ln \sigma - \frac{1}{2\sigma^2} \sum_{i \in D} (t_i - \mu)^2 + \sum_{i \in C} \ln Q\ \left(\frac{t_i-\mu}{\sigma}\right)$$

となる．ここで，$z_i = (t_i - \mu)/\sigma$としてハザード関数を式(27)で表す．

$$\lambda(z) = -\frac{d}{dz} \ln Q(z) = \frac{\phi(z)}{Q(z)} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (27)$$

対数尤度関数をμおよびσについて偏微分して0とおき$\lambda(z)$を用いて表すと，式(28)，式(29)が導かれる．

$$\sum_{i \in D} z_i + \sum_{i \in C} \lambda\ (z_i) = 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (28)$$

$$-d + \sum_{i \in D} z_i^2 + \sum_{i \in C} z_i\ \lambda\ (z_i) = 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (29)$$

ニュートン・ラフソン法やその他の計算法を用いて推計値$\hat{\mu}$および$\hat{\sigma}^2$を得ることができる．信頼区間については尤度比の分布から求めることができる[26]．

結婚市場参入の年齢Xが正規分布を持ち，参入後から結婚するまでの待ち時間Yが指数分布を持つものと仮定した場合，結婚年齢$X+Y$の分布はXとYとの分布のたたみ込みで表される[27]．結婚年齢をすべてのサンプルについて観測できないデータについても同様の手続きで尤度関数を用いてλ，μ，σを推定することが理論上可能である．

26) Lawless 前掲書（注3） 233-237 ページ参照．

27) A. J. Coale and D. R. McNeil, " The Distribution by Age of the Frequency of First Marriages in a Female Cohort", *Journal of the American Statistical Association*, Vol. 67, No. 340, 1972, pp. 743-749.

3. 比例ハザードモデル

ここではハザード率 $\lambda(t;\boldsymbol{x})$ が式(30)のように時間 t と説明変数 $\boldsymbol{x}$ の関数で表されるものとする.

$$\lambda(t;\boldsymbol{x}) = \lambda_0(t)\cdot g(\boldsymbol{x}) \quad \cdots\cdots (30)$$

$g(\boldsymbol{x})=\exp[\boldsymbol{x}\boldsymbol{\beta}]$ の場合, λ を用いて生残関数と密度関数を表すとそれぞれ式(31), 式(32)となる[28].

$$F(t;\boldsymbol{x}) = \exp\left[-\int_0^t \lambda_0(u)\,du \cdot \exp\left[\boldsymbol{x}\boldsymbol{\beta}\right]\right] \quad \cdots\cdots (31)$$

$$f(t;\boldsymbol{x}) = \lambda_0(t)\cdot \exp\left[\boldsymbol{x}\boldsymbol{\beta}\right]\cdot F(t;\boldsymbol{x}) \quad \cdots\cdots (32)$$

例えば, Tを結婚年齢, X_1 を就学期間とし, 回帰係数 $\boldsymbol{\beta}$の推計値 $\hat{\beta}_1$ が統計的に有意で負であれば, 教育水準が高ければ結婚年齢が高いことを示す. 特性 $\boldsymbol{X}_1$ と特性 $\boldsymbol{X}_2$ を持つ二人の個人を較べた場合, 両者のハザード比は

$$\frac{\lambda(t;\boldsymbol{x}_1)}{\lambda(t;\boldsymbol{x}_2)} = \frac{g(\boldsymbol{x}_1)}{g(\boldsymbol{x}_2)}$$

となり, t にかかわらず一定である. この意味で式(30)のモデルは, 通常, 比例ハザードモデルとよばれる. 式(30)において $\lambda_0(t)$ は $\boldsymbol{\beta}$の推定において任意で特定される必要のない基準ハザード関数である. それは $\boldsymbol{x}=\boldsymbol{0}$ としたときのハザード率に等しい. 関数 $g(\boldsymbol{x})$ には$\exp[\boldsymbol{x}\boldsymbol{\beta}]$ が多く用いられる. 特に, $\lambda_0(t)=\lambda$ の場合には指数回帰モデル, $\lambda_0(t)=\lambda p(\lambda t)^{p-1}$ の場合にはワイブル回帰モデルとよばれる. 両モデルとも対数変換すると説明変数が線形関数で表され, 最小値分布に従う誤差項を伴う対数線形モデルの形をしていることが示される. 関数 λ_0 が指数回帰あるいはワイブル回帰モデルで与えられるものと仮定すると, 尤度関数が定まり, 回帰係数の最尤推計値が求まる. 一方, λ_0が特定の関数で表されない場合においては, 観察される結果が比例ハザードモデルに従う事象であってもそのデータから確率分布のパラメータを特定することができず, 通常の尤度関数を求めることはできない.

通常の尤度関数を用いて回帰係数を推定するかわりに, それに代わる尤度関数がコックス等によって提唱されている[29]. それは式(33)で表される[30].

---

28) ここでは, $g(\boldsymbol{x}) = \exp\left[\boldsymbol{x}\boldsymbol{\beta}\right]$についてだけに限定している. さらに,

$$F = (F_0)^{\exp[\boldsymbol{x}\boldsymbol{\beta}]}$$

と表すと, Fの対数値の対数 $\ln(-\ln F)$ は

$$\ln(-\ln F) = \boldsymbol{x}\boldsymbol{\beta} + \ln(-\ln F_0)$$

となる. 生残確率の Product Limit推定値と t とをプロットすると, $\boldsymbol{x}$ の値のグループによってほぼ平行な曲線となる. 指数回帰モデルが該当する場合には平行な直線で近似される.

29) D. R. Cox, " Partial Likelihood", *Biometrika*, Vol. 62, 1975, pp. 269-276.
J. D. Kalbfleisch and R. L. Prentice, " Marginal Likelihoods based on Cox's Regression and Life Model", *Biometrika*, Vol. 60, 1973, pp. 267-279.

30) 式 (33) では t_i 時点において発生する事象 (死亡) は一つに限られ, 重複する場合 (同時点における複数の死亡) を含まない.

$$L(\beta)=\prod_{i=1}^{k}\frac{\exp[\boldsymbol{x}_i\ \beta]}{\sum_{l\in R(t_i)}\exp[\boldsymbol{x}_l\ \beta]} \quad \cdots\cdots (33)$$

ここで，$R(t_i)$ は時点 t_i のリスク集合（時点 t_i 直前まで生残し，時点 t_i までセンサーされないサンプルの集合）を示す．リスク集合 $R(t)$ が与えられ，死亡が時点 t で発生する場合にそれが i 番目の個人である確率は

$$\frac{\lambda(t\mid x_i)}{\sum_{l\in R(t)}\lambda(t\mid \boldsymbol{x}_l)}=\frac{\exp[\boldsymbol{x}_i\ \beta]}{\sum_{l\in R(t)}\exp[\boldsymbol{x}_l\ \beta]}$$

で与えられる．従って，式(33)は観察数 k の事象についてこのような確率の要素の積を求めた結果として「尤度」を規定している．

次に，より一般的に時点 t_i において発生する事象が重複する場合も含めた「尤度」関数は式(34)で近似される．

$$L(\beta)=\prod_{i=1}^{k}\frac{\exp[\boldsymbol{S}_i\ \beta]}{\left(\sum_{l\in R(t_i)}\exp[\boldsymbol{x}_l\ \beta]\right)^{d_i}} \quad \cdots\cdots (34)$$

ここで，d_i は時間 t_i における死亡数で，$\boldsymbol{S}_i$ は d_i 人の個人の変数 $\boldsymbol{x}$ をたし合わせたものである．つまり，D_i を時間 t_i で死亡する個人の集合とすると，$\boldsymbol{S}_i$ は $\boldsymbol{S}_i=\sum_{l\in D_i}\boldsymbol{x}_l$ と表される．すべての d_i が 1 であれば，式(34)は式(33)に等しい．対数「尤度」関数は式(35)となり，この対数「尤度」を最大にする β を求めればよい．

$$\ln L(\beta)=\sum_{i=1}^{k}\boldsymbol{S}_i\ \beta-\sum_{i=1}^{k}d_i\ln\left(\sum_{l\in R(t_i)}\exp\left[\boldsymbol{x}_i\ \beta\right]\right) \quad \cdots\cdots (35)$$

各回帰係数についての偏微分は式(36)で与えられる．正規分布の場合と同様に推定値 $\hat{\beta}$ をニュートン・ラフソン法等で求めることができる[31]．

$$\frac{d\ln L(\beta)}{d\beta_r}=\sum_{i=1}^{k}\left[S_{ir}-d_i\frac{\sum_{l\in R(t_i)}x_{lr}\exp\left[\boldsymbol{x}_l\beta\right]}{\sum_{l\in R(t_i)}\exp\left[\boldsymbol{x}_l\ \beta\right]}\right] \quad \cdots\cdots (36)$$

4. 死因別ハザード関数（Multiple Decrement 生命表の基礎）

最後に，事象が生じる原因（以下では死因と限定する）が複数存在する場合にハザード関数がどの

31）推定値 β の標準誤差はフィッシャーの情報マトリックスから求められる．その他の β の仮説についての検定法も考案されている．注3）前掲書および D. Clayton and J. Cuzick, "Multivariate Generalizations of the Proportional Hazards Model", *Journal of Royal Statistical Society*, Series A, Vol. 148, Part 2, 1985, pp. 82-117 参照．

ように応用できるかを考えよう．まず，死因 j のハザード関数を式(37)で示す．これは式(4)の確率をさらに限定して，死因が j であることを $J=j$ で表したものである[32]．

$$\lambda_j(t)=\lim_{\Delta t\to 0}\frac{\Pr\left(t\leqq T<t+\Delta t,\ J=j\mid t\leqq T\right)}{\Delta t} \qquad \cdots\cdots\cdots\cdots\cdots\cdots \quad (37)$$

死因が十分に区別できるものと仮定すると m の死因をもつ人口の全体のハザード率は次式のように個々のハザード率の和となる．二つの死因が重なる場合にはそれを新たに独立した死因と考える．

$$\lambda(t)=\sum_{j=1}^{m}\lambda_j(t)$$

全体の生残確率関数は

$$F(t)=\exp\left[-\int_0^t\lambda(u)\,du\right]$$

となり，各死因についての分布密度関数は

$$f_j(t)=\lim_{\Delta t\to 0}\frac{\Pr\left(t\leqq T<t+\Delta t,\ J=j\right)}{\Delta t}=\lambda_j(t)\cdot F(t)$$

となる．この場合，尤度関数は比例ハザードモデルの場合と同様な手続きで

$$L=\prod_{i=1}^{n}\left\{\lambda_{j_i}(t_i)^{\delta_i}\cdot\prod_{j=1}^{m}\exp\left[-\int_0^{t_i}\lambda_j(u)\,du\right]\right\}$$

に比例する．ここで j_i は i 番目の個人の死因を示しており，その他の死因による死亡をデータがセンサーされたものと同様に扱っている．

一方，Product Limit 推定値は

$$\hat{G}_j(t)=\exp\left[-\int_0^t\lambda_j(u)\,du\right]$$

$$=\prod_{i\mid t>t_i,\ J=j}\frac{n_i-d_{j_i}}{n_i}$$

を与える．人口全体の生残確率の推定値は

$$\hat{F}(t)=\prod_{i\mid t>t_i}\frac{n_i-d_i}{n_i}$$

で与えられる．ここで死因 j についての生残確率は

32）より一般的に k の死因についてそれぞれ特定の生存期間 T_1，T_2……，T_k があるとし，個人の生存期間がその最小値で定まる，つまり $T=\mathrm{Min}\,(T_1, T_2, \cdots\cdots, T_k)$ と仮定した分析も展開されている．この場合，現実に生じた死亡の死因についても確率分布を仮定するため，死亡年齢とセンサー時の年齢から死因別ハザード関数を特定することは困難となる．

$$\hat{F}_j(t) = \sum_{i \mid t \leqq t_i,\ J=j} \frac{d_{j_i}}{n_i} \hat{F}(t_i)$$

によって与えられる．ここでd_{j_i} / n_iは，経験的ハザード率と考えることができる．これにより，t才以後に死因jで死亡する確率

$$\Pr(J=j \mid t \leqq T) = \frac{F_j(t)}{F(t)}$$

を計算したり，ハザード率のトレンドをグラフ化して$\lambda_j(t)$に関するパラメトリックモデルの推定式を決定する方向性を摸索することができる[33]．

Ⅲ　結びにかえて

近年，ハザード関数を用いた統計解析の例は文献のなかに数多く見うけられる．生命表を扱う人口学研究者にとってそれは対岸の火事のように感じられていたように思われる．しかし，人口事象が時間に依存する事象である以上，その分析がハザード関数を用いた統計解析からほど遠くない場所に位置することは当然のことである．さらにこれらと遠くない場所にマルコフ過程分析，Event History分析，多次元人口理論が位置する．現在，人口学研究者に要求されている課題の一つである人口事象のタイミングの分析には，これらの理論や手法が不可欠となっている．

33）競合する死因のなかで特定の死因がとり除かれた場合の生残確率を計算するためには，その特定の死因にあたるハザード率を0とし，その他の死因のハザード率を不変として計算されることが多い．しかし，より現実的な生残確率を求めるためには，死亡仮定に関する情報，特に死因間の関係や説明変数と死因との関係についての情報，が必要となる．

Hazard Function and Life Table : An Introduction to the Failure Time Analysis

Keiichiro MATSUSHITA and Hisashi INABA

Failure time analysis has become popular in demographic studies. It can be viewed as a part of regression analysis with limited dependent variables as well as a special case of event histroy analysis and multistate demography. The idea of hazard function and failure time analysis, however, is not properly introduced to nor commonly discussed by demographers in Japan.

The concept of hazard function in comparison with life tables is briefly described, where the force of mortality is interchangeable with the hazard rate. The basic idea of failure time analysis is summarized for the cases of exponential distribution, normal distribution, and proportional hazard models. Multiple decrement life table is also introduced as an example of lifetime data analysis with cause-specific hazard rates.