生存時間解析

～事象発生までの時間を扱う統計学～

東京理科大学 理学部第二部 数学科 准教授 下川（しもかわ） 朝有（あさなお）

■ はじめに：統計学とは？

近年，様々な領域・分野において，データサイエンスや AI といった言葉が飛び交っています．多くの読者が新聞やテレビで一度は耳にしたことがあるかと思いますが，これらを支えているものの 1 つとして（もしくはそのものを構成する要素の 1 つとして），統計学が大きな役割を果たしています．本稿では，統計学，その中でも特に「ある事象が発生するまでの時間」を扱う生存時間解析という分野について紹介したいと思います．

そもそも統計学とはどのような学問であり，なぜ近年急速にその重要性が取り上げられているのでしょうか？　私たちが生きているこの世界には，知りたいことがたくさんあります．例えば「近所の湖に住む魚はどのくらいの大きさをしているのだろう？」といった素朴な疑問から，「新たに開発した新型コロナウイルスのワクチンはどの程度の効果があるのだろう？」といったより現実的な問題まで，考え出したらきりがありません．さてそれではこれらの疑問は，どのように解決することが出来るでしょうか？　少し考えてみると，実はこれらの疑問に対する真の答えを知る方法はたった 1 つしかないことに気づきます．すなわち興味の対象となる集団について，そのすべての要素を正確に調べることです．先ほどの例で言うならば，「近所の湖に住む全ての魚を捕まえて体長を測る」「世界中の全ての人に対して開発したワクチンを投与して効果を観察する」といった具合です．当然そんなことは，調査の費用や時間，また倫理的な観点から不可能であることが分かります．世の中には知りたいことがたくさんありますが，真の解を得られない場合が多く存在しているようです．

真の答えは得られないことは分かりましたが，それでも何かしら妥当な答えを得たい場合，どうすれば良いでしょうか？　この問題に対処するため，昔から人々はサンプリングということを行ってきました．すなわち，興味の対象となる集団のすべての要素を調べることは不可能だが，その一部を取り出してきて調べることで，知りたいことに対する知見を得よう，という自然で現実的な対処法です．しかしサンプリングという性質から，得られる結果には常にバラつきが内包されています．例えば今日 100 匹の魚を無作為に選び調査したところ平均が 20 cm だったからといって，明日同様の調査を行った際に全く同じ 20 cm とは限りません．そのため，この結果のバラつきを考慮しながら，得られた標本の性質を調べ，知りたかった疑問に対する推測を行っていく必要があります．この興味の対象となる集団を定めサンプリングを行い，得られた標本を用いて世の中の知りたいことに対する知見を得ていく一連の流れ全てを「統計」と呼びます．そして統計を行うには様々な疑問がついて回ります．例えば「どのようにサンプリングするべきか？」「どの程度データを集めるべきか？」「どのように推測を行い，得られた結果の正確さは？」など，知りたいことや興味の対象となる集団が変われば，疑問は形を変えて次々と現れてきます．これらの疑問に対し，数学，特に確率論に基づき理論を構築し，時にはシミュレーションを用いて考察していく学問が「統計学」となります．

それではなぜ最近盛んに統計学の重要さが取り上げられるのでしょうか？　この主要な要因の 1 つとして，データに関わる技術の急速な進歩があげられます．これは主に，データを収集するセンサー技術，データを蓄えておくためのストレージ技術，そしてデータを処理するための計算機の発展によります．読者の皆様も，身の回りのカメラやスマートフォン，PC について振り返ってみると，たった数年で急速にその性能が発展していることに驚かれるのではないでしょうか．これらの技術の進歩に伴い，幅広い分野において多くのデータが収集され，利用することが可能となっています．そしてそれらのデータの収集と分析において理論的根拠を紐づける統計学は，まさに多くの分野において必要とされる学問となっているのです．

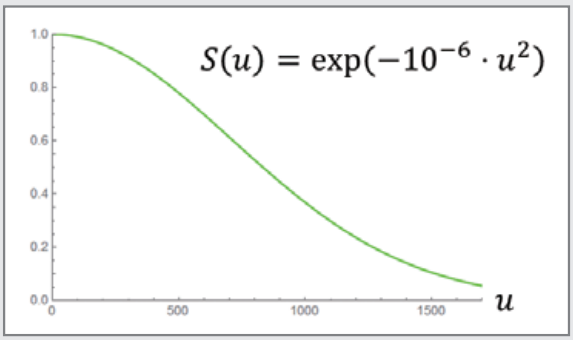
■ 生存時間　～事象発生までの時間を扱う～

ここからは統計学の中でも特に生物統計学と呼ばれる分野において広く用いられている，生存時間解析について紹介していきたいと思います．生存時間とは「明確な起点から，興味のある特定の事象が発生するまでの時間」を指します．例えば，特定の発がん性物質をマウスに投与しその影響を調べる実験における生存時間としては，「マウスに物質を投与してから死亡するまでの時間」が考えられます．また，進行性がん患者の手術後における院内感染を調べる研究においては，「手術後から感染が発生するまでの時間」が考えられるでしょう．生物分野に限る必要は無く，例えば工場で作られた電球が切れるまでの時間を調べる場合には，「電球を使いだしてから電球が切れるまでの時間」が生存時間となります．ここでは「ある病気の患者について，治療開始から死亡までの時間は？」という例を参考に，基本的な生存時間解析の話について数式を交えて紹介してみたいと思います．

統計学では興味の対象となる研究対象の集団のことを母集団と呼びます．ここでの母集団は「治療を行ったその病気の全ての患者」となります．母集団内の全ての患者は「治療開始から死亡までの時間」を必ず持っている，すなわち，観察を永遠に続けていけば，全ての患者はどこかの時点で必ず死亡する，という前提で考えていきたいと思います．

今，対象となる母集団から無作為にサンプリングを行い，得られた患者の生存時間（すなわち，治療開始から死亡までの時間（単位：日））を変数 U として表すとします．例えばサンプリングの結果得られた患者の生存時間が 50 日だとしたら，$U=50$ となります．このように標本のとりうる値に対して実数値を定める写像のことを確率変数と呼び，統計学では基本的な変数となります．

さてこの確率変数ですが，先にも述べた通り，サンプリングの性質から得られる結果はバラつきます．例えば，たまたま今回のサンプリングでは $U=50$ でしたが，もう一度サンプリングを行ったら $U=100$ かもしれません．すなわち，母集団内の各患者は異なる生存時間を持っており，その値は分布していると考えることが出来ます．この分布のことを，母集団分布と呼び，サンプリングの結果得られた標本に基づいて，この母集団分布に対する推測を行っていくことになります．ここで母集団分布を考えていくうえで，生存関数と呼ばれる関数を定義したいと思います：



【図 1】 ワイブル分布の生存関数

$$S(u)=P(U>u).$$

これは確率変数 U がある値 u より大きい値となる確率を表している関数であり，名前の通り，ある時点 u まで生存する（死亡しない）確率を表す関数となります．生存関数の例として，**【図 1】** にワイブル分布と呼ばれる分布を描いてみました．生存関数は時点 0 において 1 をとり（誰も死亡していない），時間の経過と共に単調に減少していく（死亡した人が増えていく）関数となります．

それでは標本からこのような生存関数（母集団分布）を推測することで，どのようなことが分かるのでしょうか？　実はこの生存関数は，世の中で広く用いられています．実際に読者の皆様が一度は耳にしたことがあると思われる例を幾つかあげてみたいと思います．

まず単純に，この関数のある値に着目した数値として生存率があります．例えば「3 年生存率」と呼ばれる値がありますが，これは 3 年＝1095 日を超えて生存できる確率，すなわち 3 年まで死亡しない確率として，$S(1095)$ の値が直接対応します．【図 1】の例でいうと，$S(1095)=\exp(-10^{-6}\cdot 1095^2)\approx 0.30$，すなわち約 30% となります．

続いて平均余命があります．これはある時点まで生存した人が，その後どの程度生存することが期待されるか（平均的に生存できるか）を表しており，以下の式で与えられます：

$$\mathrm{mrl}(u)=E(U-u\,|\,U>u)=\frac{\int_u^\infty S(t)\,dt}{S(u)}.$$

例えば【図 1】の関数について考えると，$\mathrm{mrl}(365)=613$ であり，すなわち，「治療から 1 年（365 日）生存できた場合，その後 613 日間の生存が期待される」ことを表しています．またこの平均余命を時点 0 で評価した値 $\mathrm{mrl}(0)$ が平均寿命であり，【図 1】の例でいえば $\mathrm{mrl}(0)=886$，すなわち「治療から平均的に 886 日間生存する」ことを表しているわけです．このように我々の周りには，実際に生存関数から得ら

れた値が広く用いられています．

■ 生存時間解析　～生存関数の統計的推測～

ここからは生存時間関数の推定の話をみていきたいと思います．今，対象となる母集団から無作為にサンプリングを行い，得られた患者は全部で n 人いると仮定し，i 番目の患者の生存時間を U_i と表記したいと思います $(i=1,2,\cdots,n)$．これらの標本 $U_1,U_2,\cdots,U_n$ を用いて生存関数の推定を行っていくわけですが，ここではパラメトリックな手法というものについて紹介したいと思います．

パラメトリックな手法は，母集団分布の形について一定の制限をかける方法となります．具体的に，先ほどの【図 1】でみたワイブル分布を参考に考えていきたいと思います．【図 1】のワイブル分布ですが，その生存関数は $S(u)=\exp(-10^{-6}\cdot u^2)$ となっていました．ここでこの関数について一般化し，$S(u)=\exp(-\lambda u^{\alpha})$ としてみたいと思います．この λ と α は分布のパラメータと呼ばれ，この値が具体的に定まることにより，生存関数の形が一意に定まることになります（例えば【図 1】の例では $\lambda=10^{-6},\alpha=2$ となります）．実際に幾つかのパラメータの値についてワイブル分布の生存関数を描いてみたのが**【図 2】**となります．パラメータの値を変えることで，様々な形の生存関数を表現可能なことが分かりますが，パラメトリックな手法では，$U_1,U_2,\cdots,U_n$ を用いていかにしてこのパラメータの値を推測していくのかが主要な課題となります．

今，母集団分布がワイブル分布に従っていると仮定した場合，標本 $U_1,U_2,\cdots,U_n$ は $S(u)=\exp(-\lambda u^{\alpha})$ に従う分布からランダムに得られているはずです．そこでサンプリングで得られた $U_1,U_2,\cdots,U_n$ の具体的な観測値が $u_1,u_2,\cdots,u_n$ であったとして，それらの値が同時に観測される確率を求めてみたいと思います．無作為なサンプリングということは，$U_1,U_2,\cdots,U_n$ は互いに独立，すなわち $U_i=u_i$ であることは別な $U_j=u_j$ となることの確率に影響を与えないため，

$$P(U_1=u_1,U_2=u_2,\cdots,U_n=u_n)=\prod_{i=1}^{n}P(U_i=u_i)$$

を満たします．

つづいて $P(U_i=u_i)$ を求めたいのですが，これを具体的に求めることは難しそうです．というのも，確率変数 U は今，生存時間を表していました．「時間」について考えてみると，その取りうる値はどのくらい存在しているでしょうか？　例えば 0 から 1 秒の間という短い時間ですら，とりうる値は無数にありそうです（1 つの値だけ考えてみても，0.251…秒と小数点以下が無限に続きます）．このように確率変数が非常に密な無限個の値をとりうる場合，その確率変数は連続型と呼ばれ，具体的にある 1 つの値をとる確率は常に 0 となってしまいます．そこで次のような関数を考えたいと思います：

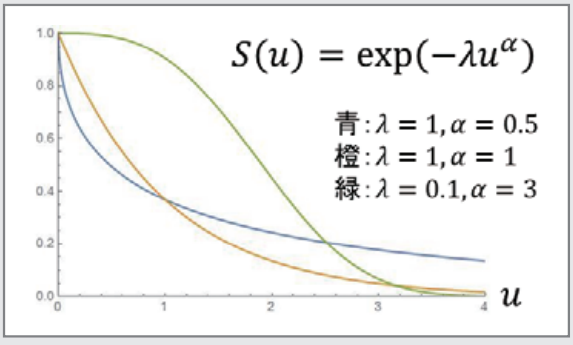
$$P(a\leq U\leq b)=\int_a^b f(x)\,dx.$$

この関数 $f(x)$ は確率密度関数と呼ばれ，確率変数がある区間に含まれる確率を，積分を用いて表現可能な関数となっています．この関数を用いることで，生存関数は $S(u)=\int_u^{\infty}f(x)\,dx$ と表現することが出来そうです．また微分と積分の性質から，$S(u)$ が微分可能な場合は，以下を満たします：

$$f(u)=-\frac{d}{du}S(u).$$

すなわち，生存関数が分かっていれば確率密度関数も定まり，例えばワイブル分布の場合は $f(u)=\alpha\lambda u^{\alpha-1}\exp(-\lambda u^{\alpha})$ となります．さて，この確率密度関数を用いて，$P(u\leq U\leq u+du)=\int_u^{u+du}f(x)\,dx$ という確率について考えてみたいと思います．もし du が十分に小さければ，$f(u)\,du$ はこの確率の近似値としてみなすことができ，$\prod_{i=1}^{n}P(U_i=u_i)$ の近似値は，$\prod_{i=1}^{n}f(u_i)\,du$ となります．

さて，$\prod_{i=1}^{n}f(u_i)\,du$ の中にはパラメータ α と λ が含まれており，これらの値が定まればこの具体的な確率が求まります．そこでこの確率を α と λ の関数とみなし，観測値 $u_1,u_2,\cdots,u_n$ が得られる確率を最大にする α と λ は何だろう？　という問題として考えてみたいと思います．すなわち「（今回の観測が得られる）最も（確率的に）尤もらしいパラメータの推定量」を考えるわけです．ここで du は定数であり，α と λ の推定には影響しないため，省いて考えてみたいと思います：



【図 2】 様々な値のパラメータに対する，ワイブル分布の生存関数

$$L(\alpha,\lambda)=\prod_{i=1}^{n}f(u_i)=\prod_{i=1}^{n}\alpha\lambda u_i^{\alpha-1}\exp(-\lambda u_i^{\alpha}).$$

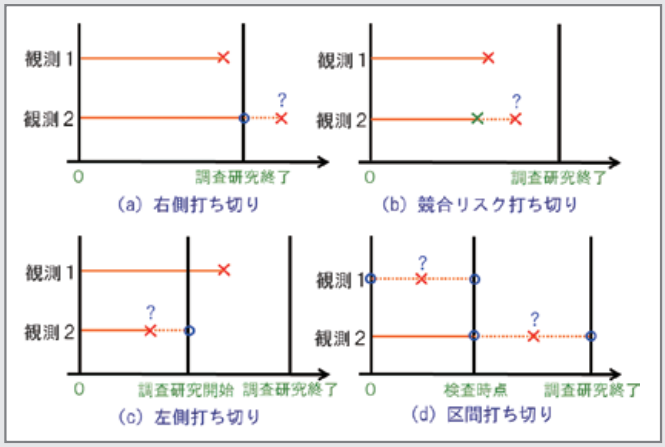
このように与えられた観測が同時に得られる確率を考え，パラメータの関数としてみたものを尤度関数と呼び，これを最大にするパラメータの推定量のことを「最尤推定量」と呼びます．この「最も尤もらしい推定量」は，幾つかの望ましい性質を持ち，統計学における主要な推定の 1 つとなっています．

■ 打ち切られた観測　〜欠損した情報を扱う〜

ここまでの話は，統計学の観点から見ればよくある解析と変わりません．すなわち生存時間解析という特定の一分野として扱わずとも，一般的な統計手法を用いることで十分解析可能です．しかし実際は，生存時間特有の問題が幾つか存在します．ここではその中でも代表的な，データの特殊な欠損構造「打ち切り」について紹介したいと思います．

打ち切りとは「事象の発生の正確な時点が分からず，ある期間内に発生したことだけが観測される」ことを言います．例えば「手術後から感染が発生するまでの時間を調べる研究」を行っているとしましょう．時間を扱う調査研究の場合，一般にその調査研究の終了時期が定まっています（もし定まっていないとすると，調査にかかる時間や費用の見積もりが立たず，研究計画が立ちません）．調査研究の終了時期が来たとき，その時点で感染が発生していない患者については正確な生存時間が分からず，「調査終了後のいつかは感染が発生するが，調査期間内では発生しなかった」ことのみが観測されます．これを表しているのが**【図 3(a)】**となります．この図では横軸が時間を表しており，赤いバツ印が着目する事象の発生時点を表しています．図内の観測 1 は調査研究開始から終了までの期間内に事象が発生しており，正確な生存時間が分かります．一方，観測 2 では青い丸印まで生存したことが分かるだけで，正確な生存時間が分かりません．このような観測を「右側打ち切り」をうけたといいます．

また「競合リスク打ち切り」もよく発生します．これは着目する事象とは異なる事象（競合リスク）が先に観測され（**【図 3(b)】**の観測 2 における緑バツ印），本来観測したかった事象の正確な発生時点が分からなくなってしまう状況を指します．例えば着目しているのは感染が発生するまでの時間でしたが，その前に病気の悪化により患者が死亡してしまったとします．このとき競合リスクは「死亡」であり，この患者に対する観測は「いつかは感染が発生したはずだが，死亡日までは発生しなかった」となります．



【図 3】打ち切りのイメージ

さて，標本にこのような打ち切られた観測が含まれている場合，それをどのように扱い生存関数の推定を行っていけばよいのでしょうか？　一番手軽な方法は「打ち切られた観測は生存時間が分からないので，標本から取り除いて解析しよう」と考えることです．すなわち観測のうち，$U_1, U_2, \cdots, U_m$ の m 件については正確に生存時間が観測されており，$U_{m+1}, U_{m+2}\cdots, U_n$ の $n-m$ 件については打ち切られ正確な生存時間が分からないとすると，$L(\alpha,\lambda)=\prod_{i=1}^{m}f(u_i)$ を尤度関数とし最尤推定量を求める方法になります．これは妥当に思えますが，残念ながら誤った結論を導く危険性があります．その原因は，打ち切られた観測が持つ情報をまったく無視してしまっていることにあります．

話を具体的にするため，$U_{m+1}, U_{m+2}\cdots, U_n$ は時点 v で右側打ち切りを受けたとしましょう．すなわち調査研究の開始から終了までの全期間が v であったとします．このとき打ち切られた観測が持つ生存時間に関する情報は「時点 v 以降のどこかで事象が発生する」なので，$U_j>v$ となります（$j=m+1, m+2, \cdots, n$）．これを踏まえて標本から得られた情報をまとめると，

$$P(U_1=u_1,\cdots,U_m=u_m,U_{m+1}>v,\cdots,U_n>v)=\prod_{i=1}^{m}P(U_i=u_i)\prod_{j=m+1}^{n}P(U_j>v)$$

となります．ここで $P(U_j>v)=S(v)=\exp(-\lambda v^{\alpha})$ であるため，先ほどと同様の議論から尤度関数は，

$$L(\alpha,\lambda)=\prod_{i=1}^{m}\alpha\lambda u_i^{\alpha-1}\exp(-\lambda u_i^{\alpha})\{\exp(-\lambda v^{\alpha})\}^{n-m}$$

となり，α と λ の最尤推定量が得られます．

一方，標本に競合リスク打ち切りが含まれる場合はもう少し複雑になります．もし競合するリスクの発生と着目する事象の発生が依存関係にある場合，すなわち「感染までの時間」と「死亡までの時間」の間に何

らかの確率的な依存関係が存在する場合（例えば，感染症の発現が早い人は，症状の悪化による死亡が早いなど），標本からそれらの依存関係を特定することは不可能であるという問題が起こります．このような場合における解析手法については盛んに研究が行われていますが，ここでは「競合リスクの発生と着目する事象の発生は独立である」と仮定できる場合についてみたいと思います．この仮定の下，j 番目の観測が時点 v_j で競合リスク打ち切りをうけたとすると，その観測が持つ生存時間に関する情報は「時点 v_j 以降のどこかで事象が発生する」なので $U_j > v_j$ となります．したがって先ほどと同様の議論から，尤度関数は

$$L(\alpha,\lambda)=\prod_{i=1}^{m}\alpha\lambda u_i^{\alpha-1}\exp(-\lambda u_i^{\alpha})\prod_{j=m+1}^{n}\exp(-\lambda v_j^{\alpha})$$

として求めることができます（正確には，競合リスクの確率構造も含める必要がありますが，α と λ の推定には影響がないため，ここでは省略しています）．

他にも，生存時間を扱う場合は様々な打ち切りが発生します．例えば「幼児がある課題をいつ習得できるかの調査研究」では，研究開始時点ですでにその課題を習得している幼児が含まれているかもしれません．これは「左側打ち切り」とよばれ，**【図 3(c)】**の観測 2 で描かれる状況になります．また「定期診断によるがんの再発までの時間の調査研究」では，前回の診断日から今回の診断日までの間のどこで再発が発生したか分かりません．これは「区間打ち切り」と呼ばれ，**【図 3(d)】**の観測 1 や観測 2 の状況となります．

生存時間を扱う際は多くの場合，何かしらの欠損が標本に含まれ，それを無視して解析すると誤った結果を導いてしまいます．どのようにその構造を考え，得られた結果を解釈するかが重要となるわけです．

■ 共変量　～不随する情報を扱う～

実際の解析では，観測された生存時間（もしくは打ち切り時間）に加え，それに影響を与える可能性を持つ複数の因子が付随してくる場合が多く存在します．例えば医療データでは，患者の性別や年齢といった人口学的情報，喫煙歴や飲酒歴といった習慣的情報，BMI や血圧といった生理的情報に加え，どのような処置を行ったかという治療情報を一般に含みます．そして多くの場合，それら（共変量やリスク因子と呼ばれます）が生存時間に与える影響を調べることで，治療法の評価や治療方針の決定に役立てることができます．

例えば今，ある病気に対する治療法 A と治療法 B が存在し，どちらの方がより優れているかに興味があるとします．「治療開始から死亡までの時間」を主要な評価基準としたとき，どのようにしてそれらを評価し比較すれば良いでしょうか？　直観的な方法としては，各治療法を受けた被験者のデータを用いて，それぞれの生存関数を推定し比較することです．しかしその場合，一方の治療法の推定された生存関数が，「全ての時点で」他方の生存関数より優れていない限り，どちらの治療法が良いかを判断することは難しそうです．例えば，各治療法について推定された生存関数が【図 2】の青と橙の曲線だった場合，どちらの方が優れていると言えるでしょうか？　前半の時点では橙の生存確率の方が高いですが，後半では逆転し，青の方が優れていそうです．

また他の共変量の影響という問題も存在します．各治療法を受けた 2 群間に含まれる被験者は互いに異なるのが一般的であり，その違いが生存時間に与える影響を調整しない限り，正しく治療法を評価し，比較することができません．例えば，肺がん患者に対して治療法 A と B を比較したいとき，治療法 A を受けた被験者内の喫煙者の方が治療法 B よりも多かったとしたら，推定された生存関数の違いは治療法によるものなのか，それとも喫煙歴によるものなのか判断がつきません．

それではこれらの問題に対し，どのようにして共変量が生存時間に与える影響を調べることができるでしょうか？　ここではよく用いられる「乗法的ハザード比モデル」を紹介したいと思います．

初めに「ハザード関数」を定義したいと思います：

$$h(u)=\lim_{\Delta u\to 0}\frac{P(u\le U<u+\Delta u\,|\,U\ge u)}{\Delta u}.$$

確率密度関数の話と同様に，この関数から $h(u)\Delta u$ は確率 $P(u\le U<u+\Delta u\,|\,U\ge u)$ の近似値とみなすことができます．すなわち Δu が十分に小さいとすると，これは「ある時点 u まで事象を経験していない対象が，次の瞬間に事象を経験する近似確率」とみなすことができ，任意の時点間における事象発生の相対的な比較ができます．また確率の計算から，

$$h(u)=-\frac{d}{du}\log S(u)$$

という生存関数との関係を導くこともでき，例えばワイブル分布の場合は $h(u)=\alpha\lambda u^{\alpha-1}$ となります．具体的に，【図 2】で示した各パラメータ値のワイブル分布に対するハザード関数は**【図 4】**となります．青い線を見てみると減少していくハザード関数となってお

り，これは時間の経過と共に事象の発生する確率が減少していくことを表しています．例えばこれは，手術直後に死亡のリスクが高い移植患者のデータを扱う場合等にみられます．また緑の線では時間の経過に対してハザードが増加していき，これは部品の摩耗による故障発生のデータを扱う場合等によくみかけます．

それでは，乗法的ハザード比モデルとはどのようなものでしょうか？　今，共変量が全部で p 個存在し，それらが $z_1, z_2, \cdots, z_p$ と表されているとしたとき，このモデルは以下の式で与えられます：

$$h(u|z_1, \cdots, z_p) = h_0(u)\exp(\beta_1 z_1 + \cdots + \beta_p z_p).$$

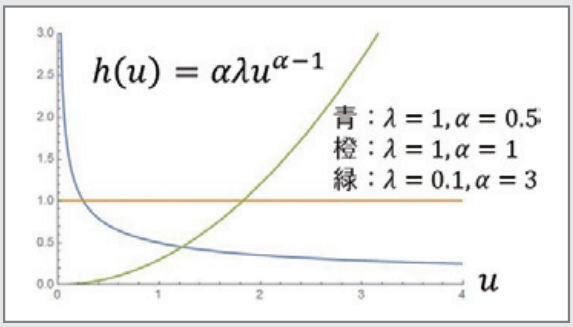
ここで $h_0(u)$ はベースラインハザードと呼ばれ，例えばワイブル分布のハザード関数を与えることができます．$\beta_1, \beta_2, \cdots, \beta_p$ は各共変量に対する係数であり，共変量の値が変化した際にハザードに与える影響度合いを表しています．

このモデルの特徴として，異なる共変量の値を持つ対象間のハザードの違いは，時間に依存しない $\exp(\beta_1 z_1 + \cdots + \beta_p z_p)$ にのみ依存する点があげられます．また，一方の対象が共変量 $z_1, z_2, \cdots, z_p$ を持ち，別な対象が共変量 $z_1', z_2', \cdots, z_p'$ を持っているとしたとき，そのハザード関数の比は以下で与えられます：

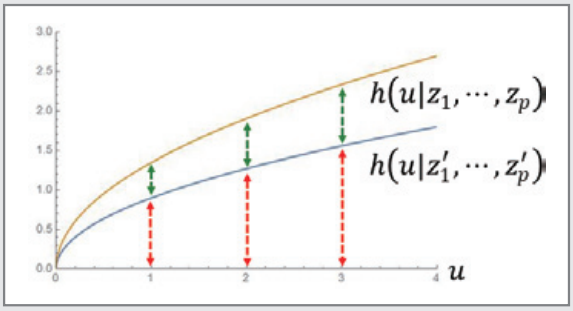
$$\frac{h(u|z_1, \cdots, z_p)}{h(u|z_1', \cdots, z_p')} = \exp\{\beta_1(z_1 - z_1') + \cdots + \beta_p(z_p - z_p')\}.$$

この値は時間に依存しないため，全ての時点において一定であり，1 より大きければ共変量 $z_1, z_2, \cdots, z_p$ を持つ方が，1 より小さければ共変量 $z_1', z_2', \cdots, z_p'$ を持つ方が，全ての時点でより生存確率が低いことを表します．**【図 5】**ではこの客観的なイメージを示しており，どの時点においても，2 つのハザードの比は常に一定となることが分かります．

このモデルを用いる際の主要な目的は，標本から係数 $\beta_1, \beta_2, \cdots, \beta_p$ を推定することにあり，それにより共変量が生存時間に与える影響を調べることができます．すなわち，着目している共変量に対応する係数の推定値が正で大きければ，その共変量の値が増加することでハザードが急激に増加することを表し，一方で，係数の推定値が負であれば，ハザードは減少することが分かります．また $\beta_1, \beta_2, \cdots, \beta_p$ は互いの共変量の影響を同時に考慮しながら推定されます．したがって先ほど見たように，治療 A と治療 B の比較を行いたい際，各治療を受けた被験者群の共変量の違いとその影響を調整したうえで，効果を調べることができるわけです．



【図4】様々な値のパラメータに対する，ワイブル分布のハザード関数



【図5】乗法的ハザード比モデルのイメージ

■ まとめ

本稿では，生物統計学において広く用いられている，生存時間解析について紹介してきました．生存時間は多くの領域で扱われ，その解析が必要とされる一方で，実際のサンプリングにおいては特殊な構造を持った標本が得られやすくなっています．ここで紹介してきた様々なタイプの打ち切り以外にも，例えば切断とよばれる標本の特殊な欠損構造，生存時間とは独立ではない打ち切り構造，また時間で値が変化する共変量構造といったものが存在し，そのそれぞれに対して多くの解析手法が研究されています．また統計学からの観点のみならず，機械学習を用いた予測という観点からも様々な研究が行われており，今後のさらなる発展が期待される分野となっています．

20 世紀の偉大な統計学者で，機械学習の分野にも多大な貢献を残した Leo Breiman は，「統計とは予測と解釈，そしてデータの処理を目的とするものでなければならない」と残しています．統計学は常にデータを扱うことを意識し，それを正しく処理し解釈を与えていく学問とも言えます．本稿を通じて，読者の方が少しでも統計学に興味を持つきっかけとなれば幸いです．