前回のQ＆A

Q1) 相関があっても因果関係があるとは限らないと書いてあるけど，こういう場合は相関は存在するのですか？　因果関係とのかかわりがよくわからない。

A1) 変数 X と変数 Y の間に因果関係があるとは，変数 X が変数 Y を引き起こすために必要であるような関係があるということです。Hill の条件では，因果関係があるといえるためには，X と Y の間の相関は，常に強くなくてはなりません。しかし，見かけの相関でない相関があっても，直接の因果関係がない場合はありえます。つまり，変数 X が変数 Y を引き起こすのに関与することもあるが，それは必要ではない，というような場合です。

* 今回は，予定を変更して時間が入ったデータを扱う方法を説明します。共分散分析も実際のデータ解析でよく使われるのですが，統計ソフトなしには計算しにくいのと，一般化線形モデルとしてまとめて説明できるので次回にまわすことにしました。

統計学第12回　時系列データと間隔データの扱い方

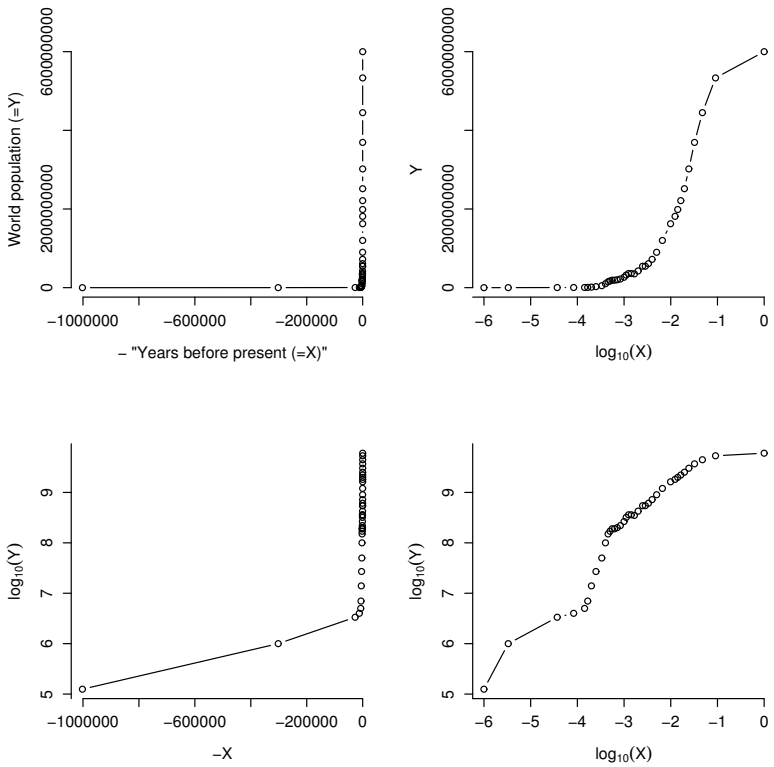
（1）時間を扱うとはどういうことか？

- 時間の入ったデータとしては，大きく分けて2種類を考えるべきである。1つは，データ間が独立でない場合である。これまで説明してきた，時間が入っていないデータは，個々のオブザーベーションが独立である。例えば，身長と体重の関係を取り上げるときは，Aさんの身長とBさんの身長には関係がないし，Aさんの体重とBさんの体重には関係がない。2次元平面にプロットされる点と点が互いに独立だからこそ，2次元正規分布に従うという仮定もできるわけである。しかし，例えば，ある人が生まれてから，毎年誕生日に身長を測って 18 歳くらいまで記録したとしたとき，年齢と身長の関係をみると，点と点の間は独立ではない。8歳のときの身長は，7歳のときの身長がどこまで伸びていたかということに，ある程度依存する。横軸に西暦年をとり，縦軸に世界人口をとってプロットした場合も同様で，1950 年の人口は，1949 年にどこまで人口が増えていたかということと無関係ではありえない。この種のデータを時系列データと呼び，時系列データを扱う解析法を総称して時系列解析という。時系列データにおける点と点の関係は微分方程式や差分方程式で表されるが，微分方程式や差分方程式を解いて非線形回帰をするよりも，自己相関をみたり，複数の波の重ね合わせとしてパタンを解析することが多い。
- 第2のパタンは，期間をデータとして扱う場合に生じる。何かのイベントが発生するまでの時間をデータとして使う場合を考えよう。例えば，結婚から第1子受胎までの時間とか，チェルノブイリで放射性物質に曝露した子どもたちが白血病を発症するまでの時間とかいったデータである。この種のデータを間隔データと総称する。時間の情報を間隔データとしてうまく使えると，少ないサンプル数でも，ある瞬間にイベントが発生する確率（ハザード）を効果的に推定することができる。出生力を推定するときに，閉経後の女性にインタビューをすることが行われるが，たんに一生のうちに子どもを何人産んだかを聞くよりも，出産暦としてすべての出産間隔を聞くほうが情報量が多いのは自明であろう。間隔データでは，観察期間中にそのイベントが起こらなかったケースは，打ち切りとなる（多くは右側打ち切り）。打ち切りデータは，イベントが起こるまでの時間がそれよりも長いケースなので，解析から取り除くと全体の推定値が過小評価されてしまう。それゆえ,「少なくとも打ち切りまでの期間より長い」という情報をうまく生かす分析法が要求される。生存時間解析とかハザード解析と呼ばれる分野で，この種の研究は多く行われてきた。
- 今回は，この2つ，即ち，時系列データと間隔データの扱い方の基礎を説明する。

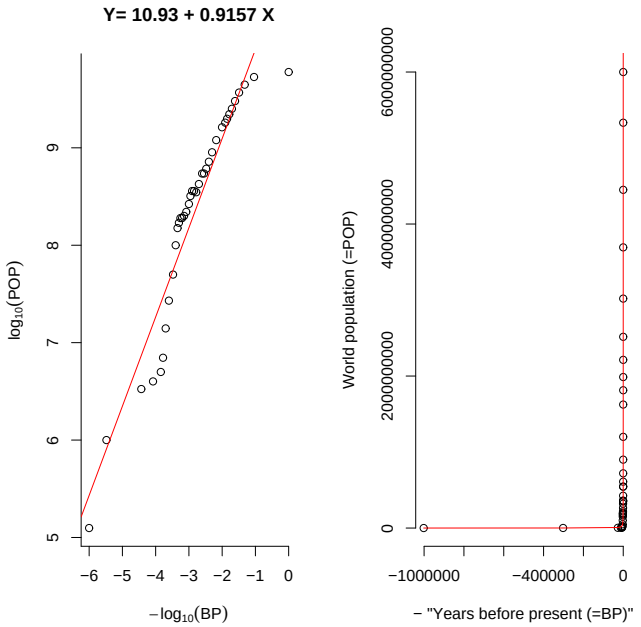
（2）時系列解析の基礎

- まず，単純な例として，さきほども取り上げた，世界人口の推移を見てみよう。ジョエル・E・コーエン著，重定南奈子・瀬野裕美・高須夫悟訳 (1998)『新「人口論」：生態学的アプローチ』（農文協刊）に載っている（米国センサス局のウェブサイトからダウンロードすることもできる）Kremer (1993) の推定値を使うと，世界人口の推移は下図のようになっている。普通の軸（左上）でみると近年の急増が激しすぎて初期の変化がわからないが，両対数でプロットすると（右下）3つの階段状（Deevey の階段，と呼ばれる）に見える[1] 。

[1] 10000 年ほど前までは，ヒトも狩猟採集に頼って生活しており，自然の生態系の一員だったと考えられるが，農耕が始まると人口支持力（環境収容力）が上がり（農耕革命），200 年ほど前の産業革命で非生物的資源を大規模に使うようになってさらに環境収容力が上がったという，不連続な人口増加率に対する説明が，一応与えられている。



- これらのデータから世界人口の将来予測をするために，変化のパタンを数学的に表そうという試みがいくつもなされてきた。最も単純なアプローチが，見かけの変化に数式を当てはめることである。前回も説明したように，直線的でない関連に対して回帰を行うには，変換して見かけ上の関連を直線に近づける方法と，非線形の数式を当てはめて最小二乗法でもっとも良く当てはまるパラメータを得る方法（非線形回帰）がある。それらの変化のパタンを使って予測することは，前回説明した回帰の外挿に当たるから，正しい保証はないし，実際，実はどれも十分な説明にならないことが既知である。ただし，もし微分方程式がメカニズム（因果関係）を正しく説明しているならば，外挿してもいいことになる（もちろん，メカニズムが変わらなければ，という限定条件の下での話しである）。だが。農耕革命や産業革命に匹敵する変化が起こったら，当然メカニズムが変わるだろうから，外挿による予測が現実に合わなくなってくるのは当然である。20 世紀後半からの少子化が，そうしたメカニズムの変化なのかどうかは，誰にもわからない。



- まず説明変数であるそのときから現在までの経過年数と，目的変数である世界人口を，両方とも対数変換する。上で示した Deevey の階段ができあがるが，これは比較的直線に近いので，とりあえず直線回帰をしてみる（左図）。対数変換していたのを元に戻したのが右の図である。かなり良く当てはまっているように見えないこともないが，左の図を見ると明らかに最近の値が過大評価になっているので，予測の信頼性はない。もっといえば，明らかに前回説明した回帰の外挿になってしまうし，そもそも，点と点が独立でないのに回帰を使うのは思想的にもおかしいから，三重の間違いである。
- 次に，微分方程式や差分方程式で隣り合う点の間の関係を説明するアプローチを説明する。これは，時系列データが互いに独立ではないことは正しく反映している。未知の係数を求めるには，非線形の最小二乗法を用いる。仮に代数的に解けない場合は，関数の最小化を数値的に行う方法がいくつか提案されているので，コンピュータを使って数値解を得る。関数の最小化を行う方法の中で最も単純なのは，Nelder-Mead の滑降シンプレックス法と呼ばれる方法である。もっと効率が良い方法としては，Powell の方法などがあるが，難解である。
- 世界人口の変化についての微分方程式モデルには，時刻を t，人口を N，増加率を r，初期人口を N_0，人口収容力を K として，
  - 指数的増加モデル: $dN/dt = rN$，即ち $N = N_0 e^{rt}$
  - ロジスティック増加モデル: $dN/dt = rN(K - N)$，即ち

$$N = \frac{K}{1 + (K/N_0 - 1)e^{-rKt}}$$

  - 最後の審判日モデル: 相対増加速度が現在人口に比例する，つまり $dN/dt = rN^2$ とするモデル。1958 年までのデータに当てはめると，2026 年には無限大に発散してしまうが，1980 年頃までは良く当てはまる（相対増加率が初めは実際より低く，後半では高すぎ）。現実の相対増加速度が減少に転じたので否定された。
  - 指数的増加の和のモデル: ２つの部分集団（先進国と途上国）に分けて，それぞれが指数的増加をすると考えたもの。先進国の割合を p として，$dN/dt = dN_1/dt + dp(N - N_1)/dt = r_1 N_1 + p r_2 (N - N_1)$ とする。現在までのところ，あてはまりは悪くない。
- これらのように微分方程式で考え，それを解いた方程式によって非線形回帰する方が[2]，変数変換によって直線に近づけ，線形回帰するよりは本質的である。人口の場合は必ず整数なので，微分方程式というよりも本質的には差分方程式であり，その意味ではカオスが生じる可能性もあるので，そもそも予測が安定しない可能性があるが，局所的には微分で考えても悪くないと仮定されている。しかし，ここに上げたモデルはどれも実際の世界人口の変化を十分に説明しきれないことがわかっている。考えてみれば，人口を構成する中身の人々は常に出生と死亡によって入れ替わり，文化も自然環境も変わっていくので，微分方程式自体がが途中で変化するかもしれないから，当てはまらないのは当然である。時系列解析の本質的な難しさは，ここにある。システムの定常性を仮定できないのである。そうなると，シナリオを仮定したシミュレーション以外には手口はない。世界人口の予測には，シミュレーション（多くは，地域を分けて出生と死亡に要因分解して各々のトレンドを予測するコウホート要因法と呼ばれる手法である）も行われている。しかし決定的な予測に成功した研究はない。
- 時系列データの予測に関しては，微分方程式や差分方程式とはまったく違うアプローチもあり，広く使われている。アプリオリに，そのデータの変化のパタンがいくつかの成分によって構成されると決めてしまい，その中身を探るアプローチである。時系列データには，繰り返し起こる（周期性がある）変化を含むように見えるものがある。例えば，日本のような中緯度に位置する場所では，一日の平均気温は，季節ごとに周期的に変化する。しかし，繰り返しは完全ではなく，地球温暖化が起これば長期的には上昇傾向をもつし，天気などによって毎日微妙に変化する。このような時系列データは，季節成分，傾向成分，不規則成分という３つの成分に分解して考えることができる。この考え方の応用としては，株価変動のような経済データから季節的な変動を除去する方法として，季節調整法と呼ばれる方法が広く用いられている。
- 周期的な変動を表す考え方の，もっとも基礎的なものは自己回帰 (Auto Regression の略で AR と呼ばれる）である。適当なタイムラグ s を置いて周期的に同じ成分によって決まる値が出現するならば，任意の時点 t における値 $x(t)$ と，時点 $t+s$ における値 $x(t+s)$ が相関をもっていることになり，

[2] R では，非線形回帰は nls というライブラリを使って実行する。`dat _ data.frame(N=POP,t=YEAR); library(nls); x _ getInitial(N ~ SSlogis(t,Asym,xmid,scal),data=dat); xx _ nls(N ~ SSlogis(t,Asym,xmid,scal),data=dat,x); summary(xx)` とすれば，N=Asym/(1+exp((xmid−t)/scal)) としたロジスティック増加モデルの係数が得られる筈だが，世界人口データではロジスティック増加モデルの当てはまりが悪いので収束せず，解が得られない。

$x(t+s)$ を目的変数，$x(t)$ を説明変数として回帰式を出したときに十分に良い fit が得られれば，s だけ後の値が予測できることになる[3] 。このとき，過去の値を十分多く使えば，予測誤差が過去の値と関係をもたなくなると期待される。式で書けば，現在の時刻を n として，時刻 n における値を示す確率変数を $x(n)$ とするとき，$x(n) = a_1x(n-1) + a_2x(n-2) + ... + a_Mx(n-M) + w(n)$ と書くと，誤差項 $w(n)$ が $x(n)$ の過去の値と独立になるということである。適切な次数 M を選ぶ方法は，期待される予測誤差の二乗が最小になるようにするのが実用的な方法の 1 つである。一般には，AIC などの情報量基準を用いて，モデルのあてはまりの悪さが最小になるようにする。$x(t)$ が 2 次定常であれば，$x(t)$ と $x(t+s)$ の共分散 $R_{xx}(s)$ は，s だけの関数となり，$R_{xx}(s) = E((x(t)-\mu)(x(t+s)-\mu))$ を自己共分散関数と呼ぶ。明らかに，$R_{xx}(0)$ は $x(t)$ の分散に等しく，$R_{xx}(-s) = R_{xx}(s)$，つまり自己共分散関数は原点について対称である。共分散の代わりに $x(t)$ と $x(t+s)$ の相関係数を考えると，$\rho_{xx}(s) = R_{xx}(s)/R_{xx}(0)$ となり，やはり s だけの関数となる。明らかに，$\rho_{xx}(0) = 1$ である。隣同士の相関が小さい確率過程の場合は，s を大きくすると急速に $\rho_{xx}(s)$ は 0 に近づく。

- 過去の時系列データから予測をすることを定式化してみよう。これは，確率過程 $x(t)$ の $s-1$ までの観測値に基づいて，次の時刻 s における値を予測することを考える。簡単のため，$x(t)$ の期待値は t によらず常に 0 とする。何らかの手段で係数の列 $\{a(m); m = 1, 2, ..., M\}$ を構成し，

$$\hat{x}(s) = \sum_{m=1}^{M} a(m)x(s-m)$$

によって $x(s)$ の予測値とすることが考えられる。このやり方で得られる予測を M 次線形予測と呼ぶ。このとき，予測誤差 $\varepsilon(s) = x(s) - \hat{x}(s)$ の 2 乗平均

$$\lim_{N \to \infty} \sum_{s=1}^{N} \varepsilon^2(s)/N$$

が最小になるように係数列 $\{a(m)\}$ を選んだときの M 次線形予測を M 次最良線形予測という。このとき，$\varepsilon(s)$ と $x(s-1), ..., x(s-M)$ は無相関となる。M を大きくしていくと，それ以上予測誤差の最小二乗平均が小さくならなくなる点が存在し，そのときの係数列を最良線形予測子と呼ぶ。このときの予測誤差の系列は，互いに相関をもたない，平均 0，分散一定の確率変数の列（すなわちホワイトノイズ）になっている。

- R で，非線形回帰以外の時系列解析をするには，まずデータを時系列データであると認識させる必要がある。例えば，ソロモン諸島ガダルカナル島で 1995 年 11 月 23 日から 12 月 29 日まで毎日朝 6 時と昼 2 時の 2 回ずつ気温を測定したデータがある。これを順番に `temp` という変数に代入するならば，`temp _ c(23.5, 28.7, 24.4, 28.5, 24.5, 30.5, 25.0, 30.9, 25.0, 26.7, 24.1, 30.3, 25.4, 28.3, 24.5, 32.9, 26.0, 29.4, 25.7, 31.2, 24.9, 29.3, 24.6, 29.9, 24.8, 32.0, 26.3, 31.8, 24.2, 31.2, 24.7, 29.6, 24.8, 30.7, 25.8, 31.7, 25.4, 30.1, 24.8, 29.1, 25.4, 30.9, 23.6, 31.2, 26.1, 30.8, 24.9, 32.2, 26.2, 31.2, 25.1, 31.9, 25.8, 32.5, 24.8, 32.4, 25.3, 31.7, 25.8, 33.6, 25.5, 31.6, 26.0, 30.5, 25.0, 33.0, 25.5, 30.0, 23.5, 31.6, 25.9, 33.0, 24.8, 33.3)` となる（本当のデータはいくつか欠損を含んでいるのだが，それだと解析が面倒なので適当に補った)。これを `ttemp` という時系列データとして認識させるには，`library(ts)` として時系列解析パッケージが使えるようにしてから，`ttemp _ ts(temp, freq=2, start=c(23, 1))` とすれば，毎日 2 回の測定値があるデータを 23 日の 1 回目から，`temp` という配列から読んで，`ttemp` という時系列データに代入する操作をしたことになる。その後の時系列解析は，この `ttemp` に対して行う。このデータに対して AR モデルを当てはめるには，`ar(ttemp)` とすればよい。この例では，4 次の自己回帰係数が計算される。

- AR 過程 $x(n)$ で表される確率システムの時刻 n における状態を $z(n)$ で表すことにすると，$z(n) = (x(n), x(n-1), ..., x(n-M+1))$ で与えられることがわかる。これと将来の入力 $w(n+1), w(n+2), ...$ がわかれば，将来の動き $x(n+1), x(n+2), ...$ が確定される。この場合，$w(n)$ は誤差項ではなく，システムを動かす入力と考えられる。すべての周波数成分を一様に含むホワイトノイズ $w(n)$ から当面の観測値の系列を生み出す確率過程の形で，多くの時系列モデルが与えられる。たとえば，$x(n)$ を 1 変量または多変量の確率過程として，状態を表すベクトル $z(n)$ を使って，$z(n) = Fz(n-1) + Gw(n)$,

[3] このためには，$x(t)$ が 2 次定常であると都合がよい。すなわち，平均と分散が時間に関して不変であり，任意の t，s に関して $E(x(t)x(s))$ が $t-s$ だけの関数となっていると都合がよい。そうでない場合は，差分を取ったり傾向成分を除去したりして，2 次定常なデータに変換することもある。

$x(n) = Hz(n)$（ここで F，G，H は，すべて適当な行列である）のように与えられる。ここで $w(n)$ を $w(n) + b_1 w(n-1) + ... + b_L w(n-L)$ で置き換えると，自己回帰移動平均過程 (ARMA モデル) が得られるが，詳細はここでは触れる余裕がないので，関心がある方は放送大学のテキストである尾崎統「時系列論」日本放送出版協会などを参照されたい。

- 次に周期的な変動をする関数としては正弦関数 (sin) と余弦関数 (cos) があるので，観察された周期性がいくつかの正弦関数と余弦関数の定数倍の和として表されると決めてしまい，この定数のセットを求めるアプローチがある（実は，すべての周期的な関数は，正弦関数と余弦関数の有限個または無限個の和で表現できることがわかっている)。物理学などで，ある光がどのような周波数の波がそれぞれどのくらいの強さで足しあわされたものかを調べる方法は，スペクトル解析 (spectral analysis) と呼ばれているが，実は周期的な波だけでなく，非周期的な波の表現方法としても有効なことがわかっており，一般にフーリエ解析 (Fourier analysis) と呼ばれる。フーリエ解析の計算方法としてよく行われるのが，高速フーリエ変換 (Fast Fourier Transformation; FFT) である。フーリエ解析を定式化すると，以下のようになる。いま，区間 $-T/2$ から $T/2$ までの間に，等間隔にとられた $2n$ 個の時系列データが与えられているとする。$x(t)$ は，$t = 0, \pm T/2n, \pm 2T/2n, ... \pm (n-1)T/2n, -T/2$ で与えられている[4] 。このとき，

$$x(t) = a_0 + 2\sum_{m=1}^{n-1}(a_m \cos\frac{2\pi mt}{T} + b_m \sin\frac{2\pi mt}{T}) + a_n \cos\frac{2\pi nt}{T}$$

と書ける。周波数 $1/T$ の整数倍の波の重ね合わせと考えるので，$1/T$ は基本周波数と呼ばれる。R では，例えば `fft(ttemp)` とすれば時系列データ `ttemp` に対して FFT が行われる。フーリエ解析については，ヒッポファミリークラブによって作られた「フーリエの冒険」という素晴らしい入門書があり，お薦めである。

- フーリエ解析では，正弦関数や余弦関数は時間と独立である。しかし，世の中には，時間とともにばらつきが大きくなるような繰り返しもある。その場合は，観察された関数を，周期的な関数で時刻を説明変数に含む，ウェーブレット (Wavelet) 関数の足し合わせに分解する方法がある。このやり方はウェーブレット解析と呼ばれ，非常に強力だが，まだ実際に使われた事例が少ない（2001 年 12 月第 1 週の Nature に，麻疹の流行パタンの分析に適用した論文が掲載されていた)。

（3）生存時間解析の基礎

- 期間データを扱う方法としては，一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれるものがある。なかでもよく知られているものが Kaplan-Meier の積・極限推定量である（現在では，普通，カプラン・マイヤ推定量と呼ばれている)。カプラン・マイヤ推定量は，イベントが起こった各時点での，イベントが起こる可能性がある人口（リスク集合）あたりのイベント発生数を 1 から引いたものを掛け合わせて得られる，ノンパラメトリックな最尤推定量である。複数の期間データ列の差の比較には，ログランク検定や一般化ウィルコクソン検定が使われる。が，ログランク検定でも Mantel-Haenzel 流のログランク検定と Peto and Peto 流のログランク検定があったり，一般化ウィルコクソン検定でも Gehan-Breslow 流と Peto-Prentice 流があったりして，非常に面倒な話になってくるので，この講義では説明しない。それらのノンパラメトリックな方法とは別に，イベントが起こるまでの時間が何らかのパラメトリックな分布に当てはまるかどうかを調べる方法もある。当てはめる分布としては指数分布やワイブル分布がある。共変量の影響をコントロールするためには，基準となる個人のハザードに対して $\exp(\sum \beta_i z_i)$ という比例定数の形で掛かるとする比例ハザード性を仮定する方法と，パラメトリックなモデルに対数線形モデルの独立変数項として入れてしまう加速モデルがある。生存時間解析も，時系列解析と同じく，それだけで一冊の本ができるほど奥が深いので，今回はカプラン・マイヤ推定量の求め方だけ説明する。より詳しくは，大橋靖雄・浜田知久馬「生存時間解析：SAS による生物統計」（東京大学出版会）などを参照されたい。
- なお，データ数が多い場合は，個々の間隔データを集計して，生命表解析を行うこともある。生命表解析の代表的なものは，ヒトの平均寿命を計算するときに行われている（官庁統計としても，まさしく生命表という形で発表されている)。平均寿命とは 0 歳平均余命のことだが，これはヒトが生まれてから平均してどれくらいの期間生存するのかという値である。一般に x 歳平均余命は，x 歳以降の延べ生存期間の総和 (T_x) を x 歳時点の個体数 (l_x) で割れば得られる。延べ生存期間の総和は，年齢別死亡率 q_x が変化しないとして，$l_x(1 - q_x/2)$ によって x 歳から $x+1$ 歳まで生きた人口 L_x（開始時点の人口が決まっていて死亡率も変化しないので x 歳の静止人口と呼ばれる）を求め，

[4] データの個数が $2n$ 個なので，区間が $-T/2$ 以上 $T/2$ 未満であると考えて，$T/2$ は入れない。

それを x 歳以降の全年齢について計算して和をとることで得られる。ヒトの人口学では年齢別死亡率から q_x を求めて生命表を計算するのが普通だが，生物一般について考えるときは，同時に生まれた複数個体（コホート）を追跡して年齢別生存数として l_x を直接求めてしまう方法（コホート生命表）とか，たんに年齢別個体数を l_x と見なしてしまう方法（静態生命表，偶然変動で高齢の個体数の方が多い場合があるので平滑化するのが普通）がよく行われる。

- では，簡単な例を使って，カプラン・マイヤ推定量を説明しよう。以下の表は，ソロモン諸島のある村で，既婚女性全員に，自分の誕生日，第１子誕生日，第２子誕生日，……，末子誕生日（まだ出産を完了していない年齢の女性も含めて，ともかくそれまでに産んだ子どもの誕生日を全部）聞き取った結果である。間隔データを使わなければ，このデータから出生力について何かいうためには，出産を完了した女性についての平均出産数（平均完結パリティという）くらいしかなくなってしまうが，間隔データを使えば，時間当たりの出生力を考えることができるので，出産を完了していない女性のデータも使うことができる。

MO_ID	MO_BD	C1_BD	C2_BD	C3_BD	C4_BD	C5_BD	C6_BD	C7_BD	C8_BD	C9_BD	C10_BD	C11_BD
20102	390000	0	640600	680000	711014	760000						
60202	250000	480415	560921	630000								
50102	400000	550000	590000	630000	660810	681011	710319	741018	760611	0		
30602	450000	580000	601004	630000	650000	670000	670000	720000	740000	750000	780714	
10502	400000	600716	630000	650807	670000	690609						
10102	400000	651103	681225	0	720200	0	790517	0	820000	840503	860527	890302
30102	490000	680000	700000	720000	750000	770000	820927					
10202	490000	680000	720826	760000	830000							
40302	580000	700000	780606	820906	901012	910606						
40102	570000	710114	730000	750000	770000	810621	840101	870802	920813			
20502	580000	720906	740704	761106	800407	811126	860516	910406				
50302	520000	730000	780000	800000	830000	870000	0					
10402	441101	730324	760723	770801	880119							
60302	460000	740000	770000	790000	800000	820000						
70202	550000	740000	780000	800000	840000	870000	890000	920000	941100			
70302	600000	750000	780000	800000	820000	850500	860000	880000	920000	940000		
20302	610000	760709	771020	790309	811002	850415	890803					
30702	600000	810500	820000	830000	840000	850000	900924	930430	950604			
30502	501205	820921	840803	881228								
60402	530000	830212	850216	900916	950921							
10802	650521	840623	861009	890727	920329	940416						
50402	670000	861114	880430	900130	910000	930325	950108					
20602	651114	870904	881111	900519	911104							
60102	570000	880000	950905									
10902	670000	900000	910000	950319								
30202	710000	900408	920210	940305								
40202	640000	901007	931109									
60204	680000	910000	920000									
50202	640000	911001	921020									
20202	711014	920801	931127									
10302	720826	920823	940308									
11002	700917	930303	950513									
10702	670304	930701										
80102	720229	940125										
11102	670809	940406										
30302	720000	940611										
50303	730000	950300										
10602	740700	950317										
20504	740704	950905										
60303	740000	951024										
70102	0	420000	450000	470000	520000	531225	550000	630000	670000			

- この種のデータには，以下の利点と欠点がある。
  - 母親に対して，全ての子どもの出生年月日を聞き取ることは，統計がしっかりしていない社会でも比較的信頼性の高い方法である。
  - 人口規模が小さくても使える上，過去の推計もできるという利点がある。
  - 古くなるほど誤差が大きくなるバイアスや，他に影響を受ける要因が多いのは欠点。
- 結婚から第１子誕生までの期間や，第１子と第２子の出生間隔がよく使われるが，上にあげたソロモン諸島の社会では，結婚記念日はあまり正確に記憶されていなかったために，第１子と第２子の出産間隔を使うことにした。第１子と第２子の出産間隔には，第２子の在胎期間が含まれるために，その期間のハザードは原理的にゼロであることに注意する必要がある[5]。
- まず，カプラン・マイヤ推定量についての一般論を示す。イベントが起こる可能性がある状態になってから，イベントが起こった時点を $t_1, t_2, ...$ とし，t_1 時点でのイベント発生率を d_1，t_2 時点でのイベント発生率を d_2，以下同様であるとする。また，時点 $t_1, t_2, ...$ の直前でのリスク集合の大きさを $n_1, n_2, ...$ で示す。リスク集合の大きさとは，その直前でまだイベントが起きていない（この例では第１子出産後で第２子出産前の）個体数である。観察途中で死亡や転居などによって打ち切りが生じるために，リスク集合の大きさはイベント発生によってだけではなく，打ち切りによっても減少す

[5] 例えば，在胎期間の推定値として 9ヶ月を引いた値をデータにしたり，または在胎期間を切片として含んだハザード関数を推定することも考慮するべきである。

る。従って n_i は，時点 t_i より前にイベント発生または打ち切りを起こした個体数を n_1 から除いた残りの数となる。なお，イベント発生と打ち切りが同時点で起きている場合は，打ち切りをイベント発生直後に起きたと見なして処理するのが慣例である。このとき，カプラン・マイヤ推定量 $\hat{S}(t)$ は，

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)... = \prod_{i<t}(1 - d_i/n_i)$$

として得られる。その標準誤差はグリーンウッドの公式により，説明は省略するが，

$$var(\hat{S}) = \hat{S}^2 \times \sum_{i<t} \frac{d_i}{n_i(n_i - d_i)}$$

で得られる。なお，カプラン・マイヤ推定量を計算するときは，階段状のプロットを同時に行うのが普通である。

- R で生存時間解析を行うには，`library(survival);` としてそのパッケージを呼び出し，`dat _ Surv(`生存時間，打ち切りフラグ`)` 関数で生存時間データを作り（打ち切りフラグは 1 でイベント発生，0 が打ち切り。ただし区間打ち切りの場合は 2 とか 3 も使う），`res _ survfit(dat)` でカプラン・マイヤ法によるメディアン生存時間が得られ，`plot(res)` とすれば階段関数が描かれる。イベント発生時点ごとの値を見るには，`summary(res)` とすればよい。
- 例えば，区間打ち切り（イベント発生までの時間がある幅をもってしかわからないデータ）を無視して，上で示したソロモン諸島のデータのうち，第 1 子出生が 1986 年以降のものの出産間隔データを R で分析すると，`time _ c(17,14,22,37,12,15,19,26,29,23,20,18,9,9,3,2); event _ c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0); dat _ Surv(time,event); res _ survfit(dat); print(res); summary(res); plot(res)` とすれば，右側打ち切りを考慮した出産間隔のメディアンが 22ヶ月であることがわかる（プロットは下図）。