数理人口学のはなし
～数理モデルを人口に適用する～

群馬大学大学院医学系研究科
生態情報学・助教授
中澤　港
<nminato@med.gunma-u.ac.jp>

数理人口学とは？

- 人口学的方法のコアにあるものは，人口にかかわる諸現象を数理モデルによって解明・理解しようとする数理人口学(mathematical demography)である……（中略）……数理人口学は人口学的概念や法則性の論理的帰結として人口現象を解明・理解することを目的としている（稲葉寿『数理人口学』東京大学出版会，2004）
- 死亡率が下がると高齢者の割合はどうなるかとか，流産の出生率への影響とかいった「常識的問題」の答えは，必ずしも常識的ではなく直感に反することもある。数学的に厳密な論理により正しい解の方向を示すことが数理人口学の目的である（抄訳，Keyfitz N, “Applied Mathematical Demography 2nd ed.”, Springer-Verlag, 1985.）

大雑把な把握

- 通常の人口分析は記述的な分析が中心。しかし要因の分析をする場合は大集団を相手にする場合が多いだけにデザインによって要因をコントロールした実験をすることはほぼ不可能なのでモデルの当てはめをすることが多くなるし，将来予測をしようと思ったらモデルの当てはめをするしかない。
- 数理人口学とは，ごく大雑把にいえば，人口現象（出生，死亡，人口成長など）に数理モデルを当てはめる方法論といえる。
- 通常，離散値を線型補間したり平滑化することは「補整」と扱われるが，モデルの当てはめであることを意識すべき。

1．死亡のモデル

ヒトの死とは？

- 誰でもいつかは必ず死ぬ
- 個人個人異なる
  - 先天異常などで生まれつき死にやすい人
  - 交通事故や戦争で死ぬ人
  - 百歳以上まで生き延びた後に老衰で死ぬ人
    ……等々
- 遺伝と環境の両方の影響を受ける
- 集団としてみれば？
  - 一般に，途上国の人は全体として先進国の人より若くして死んでしまう人が多い＝途上国は先進国より「死亡水準」が高い→どうやって表す？

ヒトの寿命

- 寿命とは？
  - ふつうの語感からすれば，事故などがなかったとき，老化が進行して機能停止に至るまで
  - なぜ老化する？
    - 体細胞廃棄説
    - テロメア説
- 死亡水準の指標
  - 乳児死亡率：途上国の衛生・栄養状態を反映
  - 平均寿命：死亡年齢の平均値ではない
- 平均寿命は年齢別死亡率を静止人口モデルで計算した値（遺伝と環境の両方含む）

ヒトは加齢に伴ってどのように死ぬか？

- 先駆的研究
  - Graunt (1662) ロンドンの人口構造を推測するのに年齢依存の死亡スケジュールを仮定
- 最初の数理的アプローチ
  - DeMoivre (1725) が生存関数l(x)を年齢xの一次関数として定式化。l(x)は，その集団のうち正確な年齢x歳まで生き延びる割合を意味する

$$\ell(x) = \ell(0)(1 - \frac{x}{\alpha})$$

α は生存する最高齢。この定式化は明らかに誤りだが，昔はそんなに悪い近似でもなかった

生命表と生存時間解析

- 生命表は生存時間解析の離散形式。
- 生存時間解析のうち，カプラン＝マイヤ法はモデルの当てはめではない。コックス回帰は比例ハザード性だけを仮定したモデルの当てはめ。加速モデルは完全に理論的な数理モデルの当てはめ。
- 通常，加速モデルでは指数分布とかワイブル分布のような単純な関数が当てはめられるが，実際の人口に当てはめるには，もっと複雑な関数であって当然。

ℓ(x)とμ(x)の様々なパタン

(Source: Denny C. *Math. Comp. Model.* 26(6): 69-78, 1997.
Vaino C. *Demog. Res.* No.3, Article 6, 2000.)

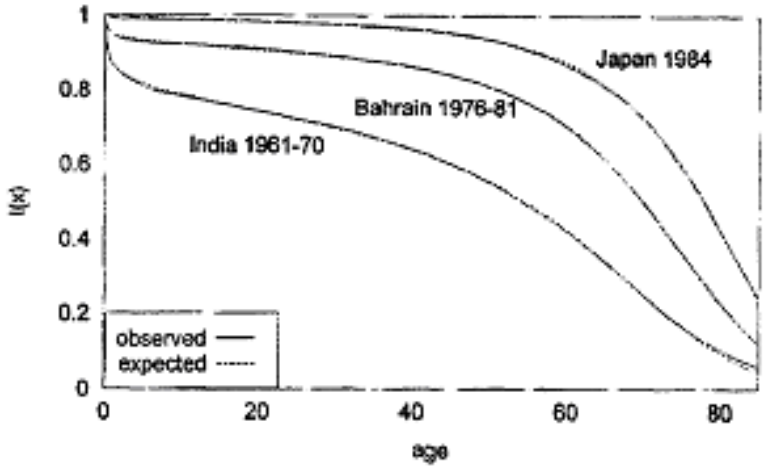


Figure 1. Observed and expected values of $\ell(x)$ for selected national male life tables.

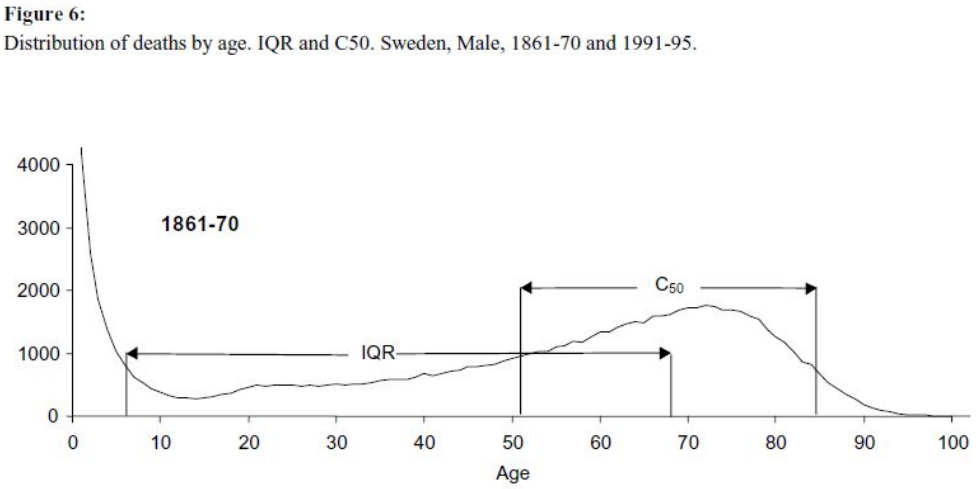
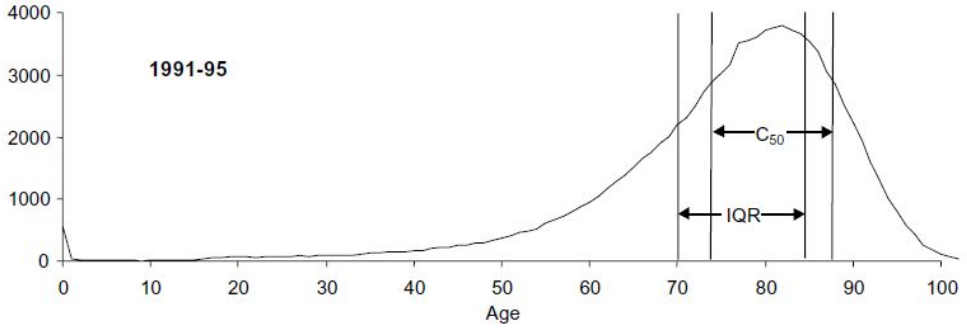


Figure 6:
Distribution of deaths by age. IQR and C50. Sweden, Male, 1861-70 and 1991-95.

$$\mu(x) = \frac{-dl(x)}{l(x)\,dx}$$

死力関数μ(x)と生存関数l(x)の関係



現実の死亡パタンへの接近

$$\textbf{Gompertz, 1825} \qquad \mu(x) = BC^x$$

$$\textbf{Makeham, 1860} \qquad \mu(x) = A + BC^x$$

$$\textbf{Perks, 1932} \qquad \mu(x) = \frac{A + BC^x}{KC^{-x} + 1 + DC^x}$$

- Gompertz (1825)は，高年齢での年齢と死力の間に指数的な関係を仮定
- Makeham (1860)は，年齢に関係ない定数項を加えた
- Perks (1932)は，さらに現実に近づくよう改変

データに基づいたモデル生命表

- 国連 (1955) 異なる死亡水準をもつ男女別の24のモデル生命表
- Coale and Demeny (1966, 1983改) 4つの異なるパタン（乳児，子供，成人，高齢者の死亡割合が異なる，東・西・南・北）と25の異なる死亡水準をもつモデル生命表
- 致命的な弱点→柔軟性がないこと。あくまで，現実のデータに「近い」モデルを選ぶしかない

Brassの方法

- Brass (1968): 数学的なモデルを，データに基づいたモデル生命表と組み合わせた
- 基準となるモデル生命表を使って，どのような生存関数も2つのパラメータで合成できると提案

$$\frac{1}{2}\ln\left(\frac{1-\ell_0(x)}{\ell_0(x)}\right) = a + \frac{b}{2}\ln\left(\frac{1-\ell_s(x)}{\ell_s(x)}\right)$$

3分割モデル
（若年，中年，老年別々に表現）

Hazard function $h(t)=\dfrac{-dN(t)}{N(t)dt}$ $(N(t)$: population at age $t)$

Thiele (1871) $h(t)=h_1(t)+h_2(t)+h_3(t)$ (h_1 for young, h_2 for middle, h_3 for elderly)

Mode (1985) $h_1(t)=\beta_1\delta_1(t+\gamma_1)^{\delta_1-1}\mathrm{e}^{\alpha_1-\beta_1(t+\gamma_1)^{\delta_1}}$, $h_2=\mathrm{e}^{\alpha_2-\beta_2(\ln\gamma t)^2}$,

$$h_3(t)=\beta_3\delta_3(t+\gamma_3)^{\delta_3-1}\mathrm{e}^{\alpha_3+\beta_3(t+\gamma_3)^{\delta_3}}$$

Gage (1991) $h(t)=\alpha_1\mathrm{e}^{-\beta_1 t}+\alpha_2+\alpha_3\mathrm{e}^{\beta_3 t}$

Probablity of death within a year at exact age $x=q(x)$ where $l(x)=l(0)\prod_0^{x-1}(1-q(x))$

Heligman and Pollard (1980) $q(x)=q_1(x)+q_2(x)+q_3(x)$

$$q_1(x)=A^{(x+B)^C},\ q_2(x)=D\,\mathrm{e}^{-E(\ln x-\ln F)^2},\ q_3(x)=\frac{GH^x}{1+GH^x}$$

Denny (1997) $l(x)=\dfrac{1}{(1+a(\frac{x}{105-x}))^3}+b\sqrt{\mathrm{e}^{\frac{x}{105-x}}-1}+c(1-\mathrm{e}^{-2x})$

- パラメータが多い（複雑な）モデルほど現実のヒトの死亡曲線によく当てはまる（Dennyは例外）

Gage, T. (1991)のモデル（再掲）

$$h(t) = \alpha_1 \cdot \exp(-\beta_1 t) + \alpha_2 + \alpha_3 \cdot \exp(\beta_3 t).$$

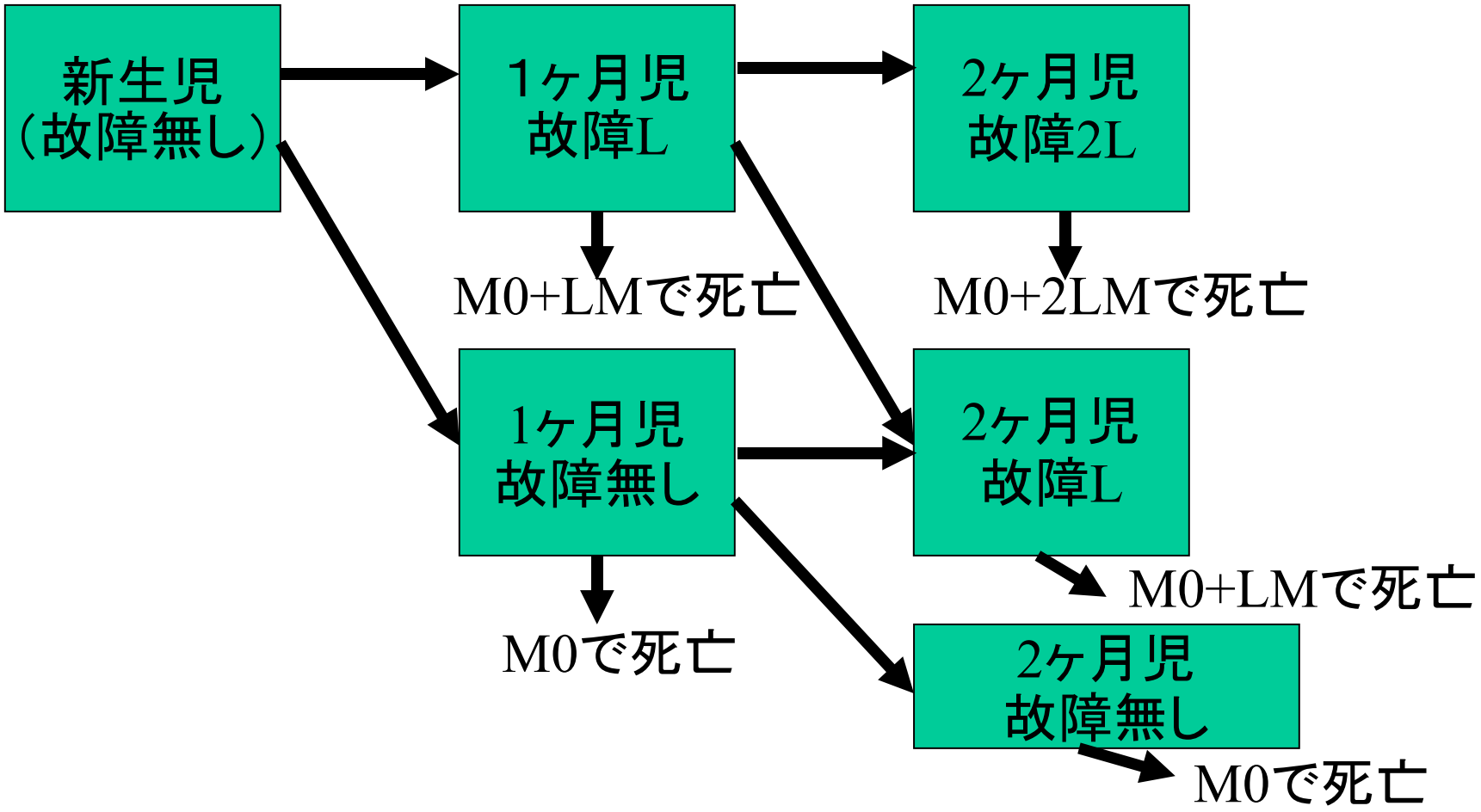
- 比較的単純で，よく現実データに適合
- 第1項は若年での感染症による死亡を意味し，第2項は事故死と妊娠出産に関連した死亡を意味し，第3項は中高年での癌や心血管疾患での死亡を意味する
- 衛生状態の改善とα1の低下が対応するなど，ある程度，パラメータが現実的意味をもつ

Gavrilov and Gavrilova (1991)

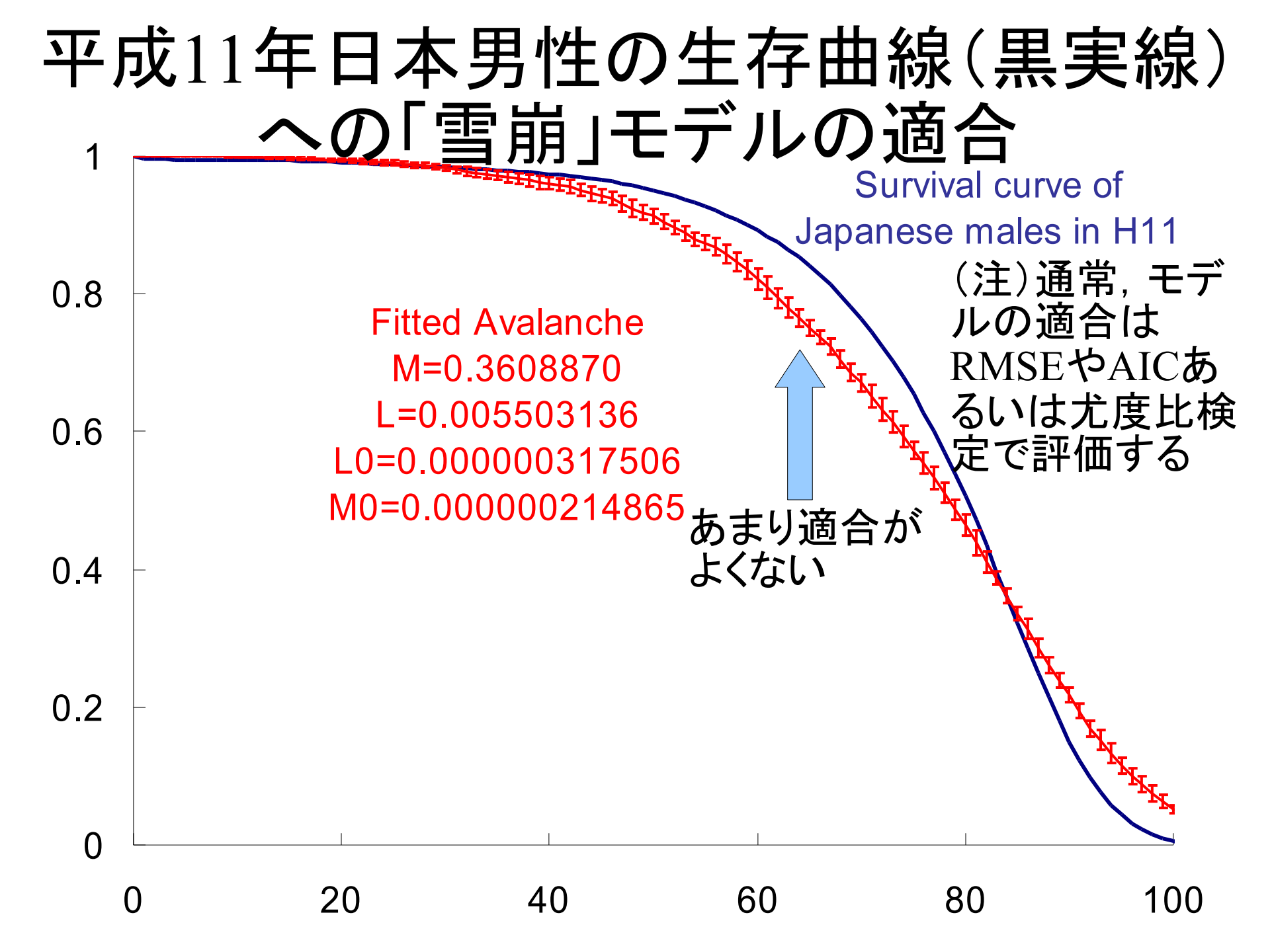
- 集団レベルの死亡曲線や生存曲線への当てはめではない，個人ベースのモデル
- 高年齢での死亡率の急増パタンから，「雪崩」モデルと呼ばれる
- 各個人はランダムにL0でバックグラウンド故障蓄積すると同時に，それまでに蓄積した故障に比例して（比例定数L）故障蓄積し，ベースライン死亡率M0に対して，それまでに蓄積した故障分だけ超過死亡（比例定数M）する（パラメータはRMSEを最小化するように探索）
- 期待値はGompertz-Makehamモデルの第2項，第3項と一致する

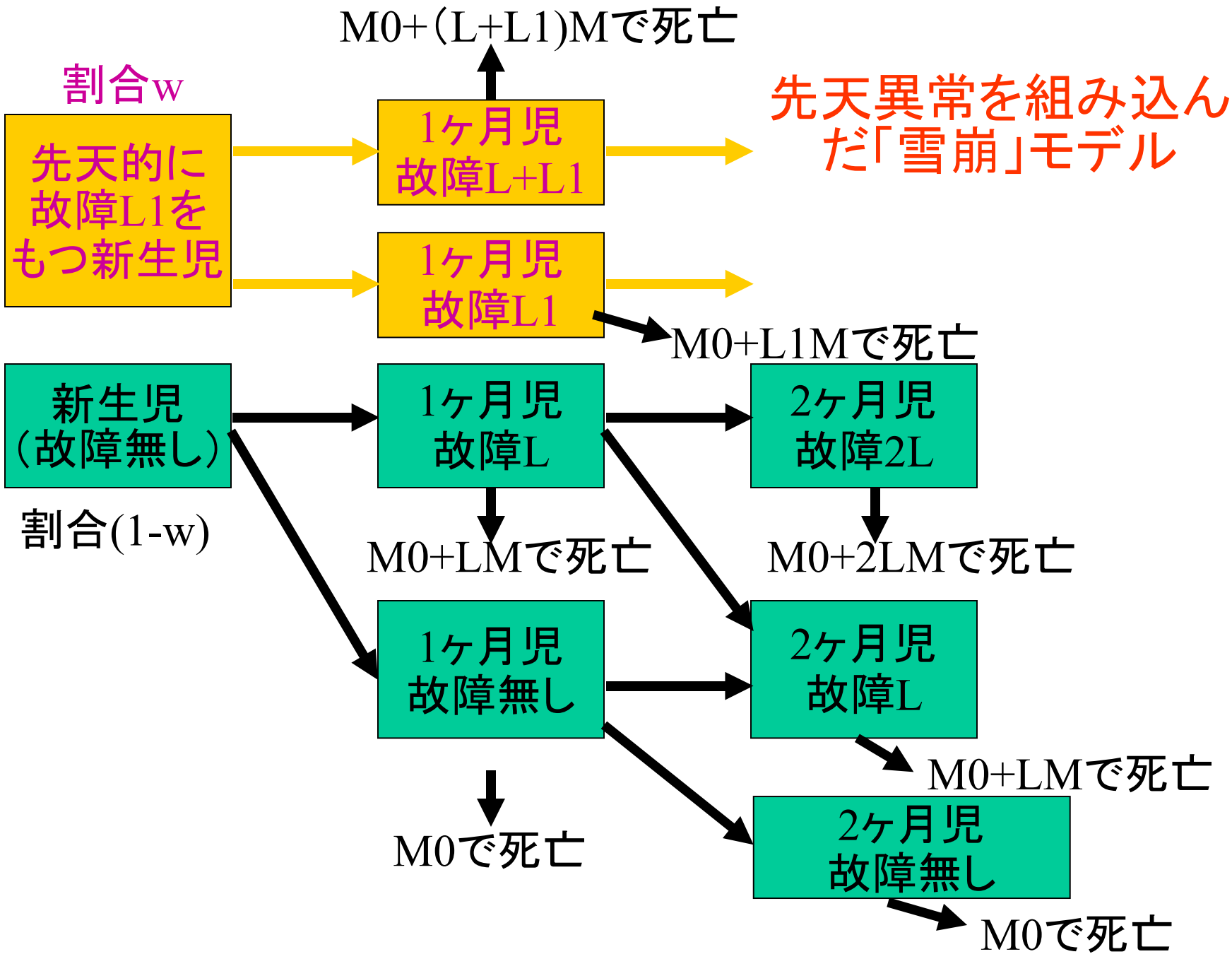
$$h(t)=\frac{\mu\lambda_0\left(1-\mathrm{e}^{-(\lambda+\mu)t}\right)}{\mu+\lambda\,\mathrm{e}^{-(\lambda+\mu)t}}$$

「雪崩」モデルのスキーム

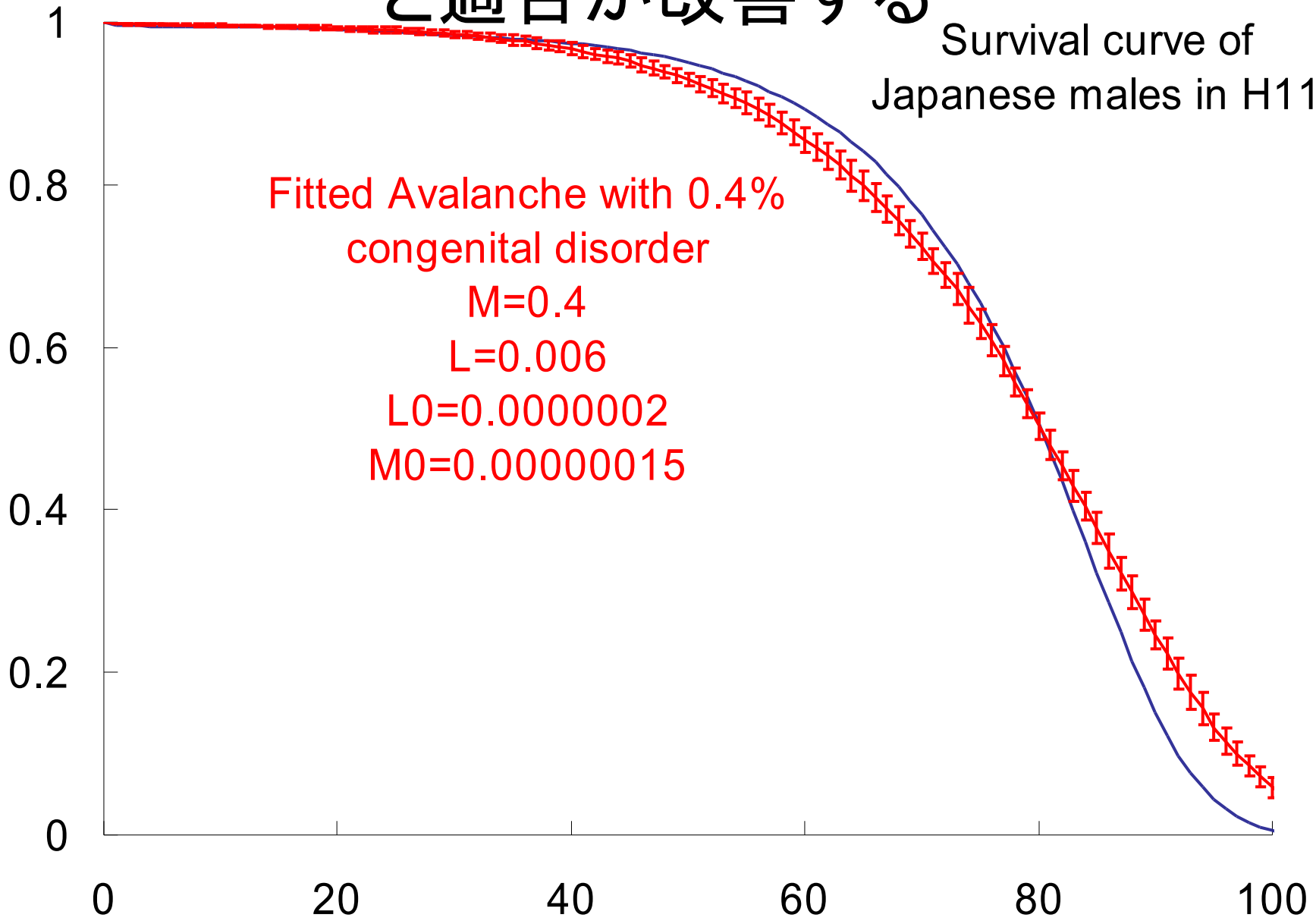

新生児
（故障無し）
1ヶ月児
故障L
2ヶ月児
故障2L
M0+LMで死亡
M0+2LMで死亡
1ヶ月児
故障無し
2ヶ月児
故障L
M0+LMで死亡
M0で死亡
2ヶ月児
故障無し
M0で死亡

平成11年日本男性の生存曲線(黒実線)への「雪崩」モデルの適合

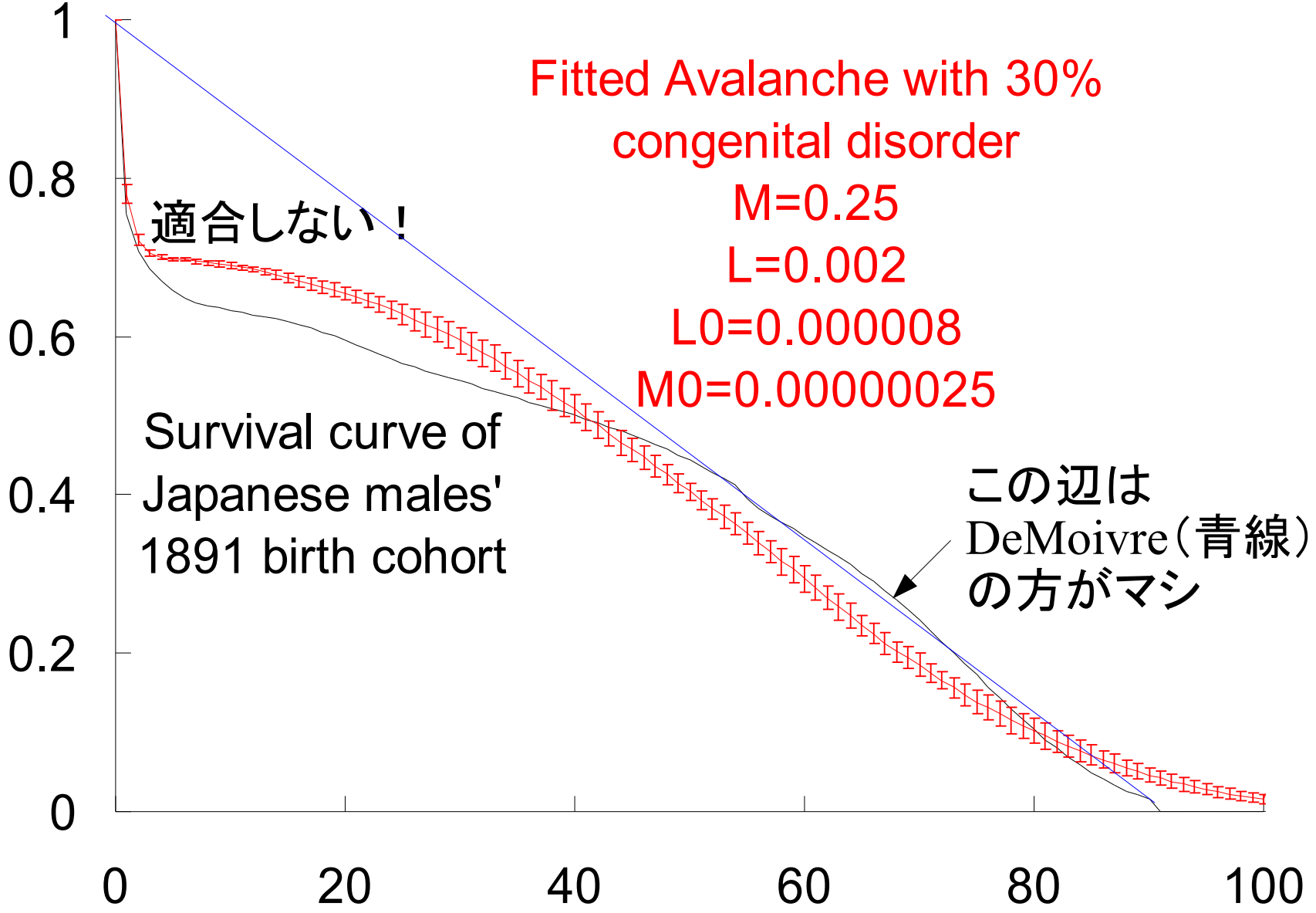



M0+(L+L1)Mで死亡
割合w
先天的に
故障L1を
もつ新生児
1ヶ月児
故障L+L1
先天異常を組み込んだ「雪崩」モデル
1ヶ月児
故障L1
M0+L1Mで死亡
新生児
(故障無し)
割合(1-w)
1ヶ月児
故障L
2ヶ月児
故障2L
M0+LMで死亡
M0+2LMで死亡
1ヶ月児
故障無し
2ヶ月児
故障L
M0+LMで死亡
M0で死亡
2ヶ月児
故障無し
M0で死亡

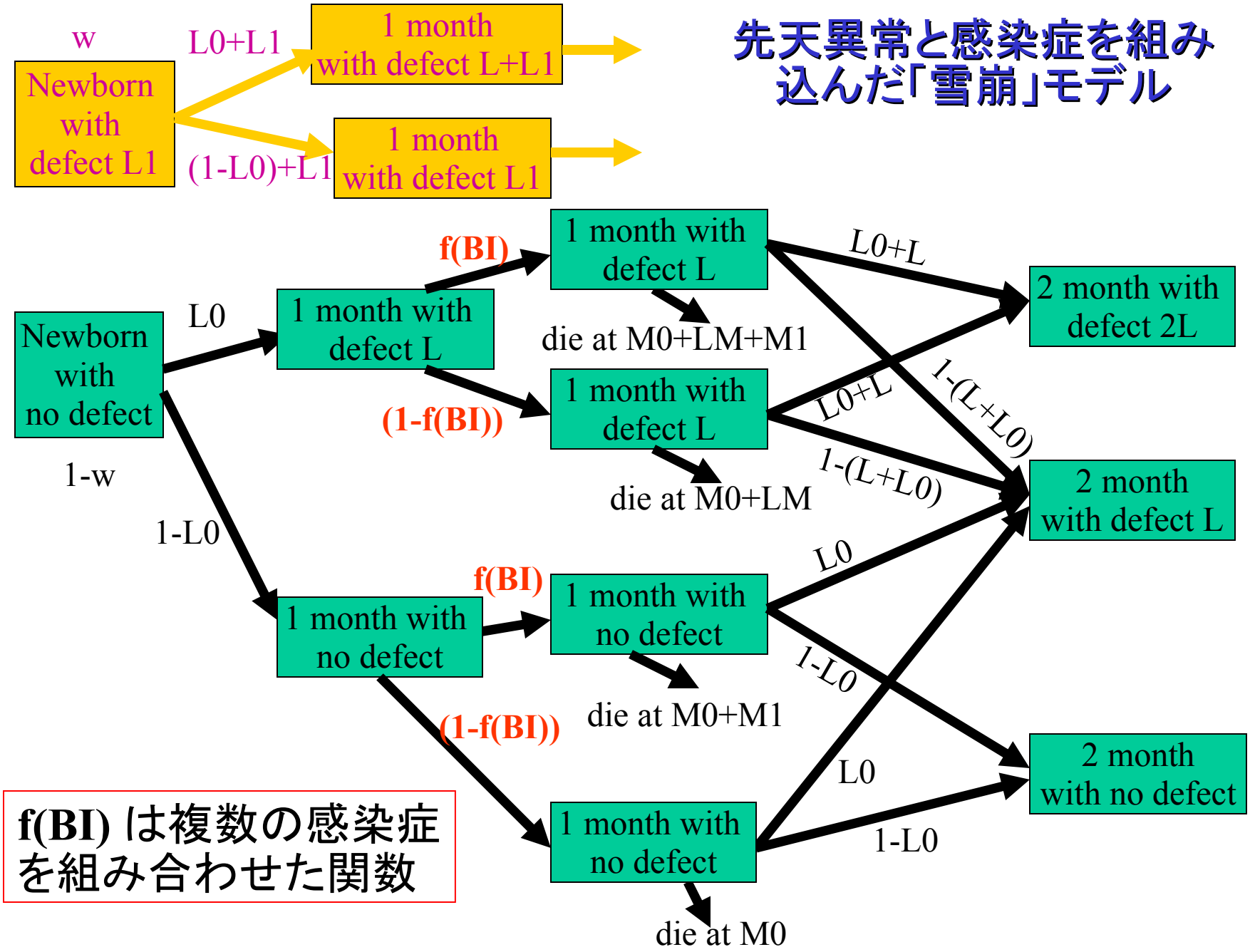
先天異常（初期故障）を0.4%組み込むと適合が改善する



1891年生まれの日本男性コホートの生存曲線には先天異常を組み込んだ「雪崩」モデルでも不適



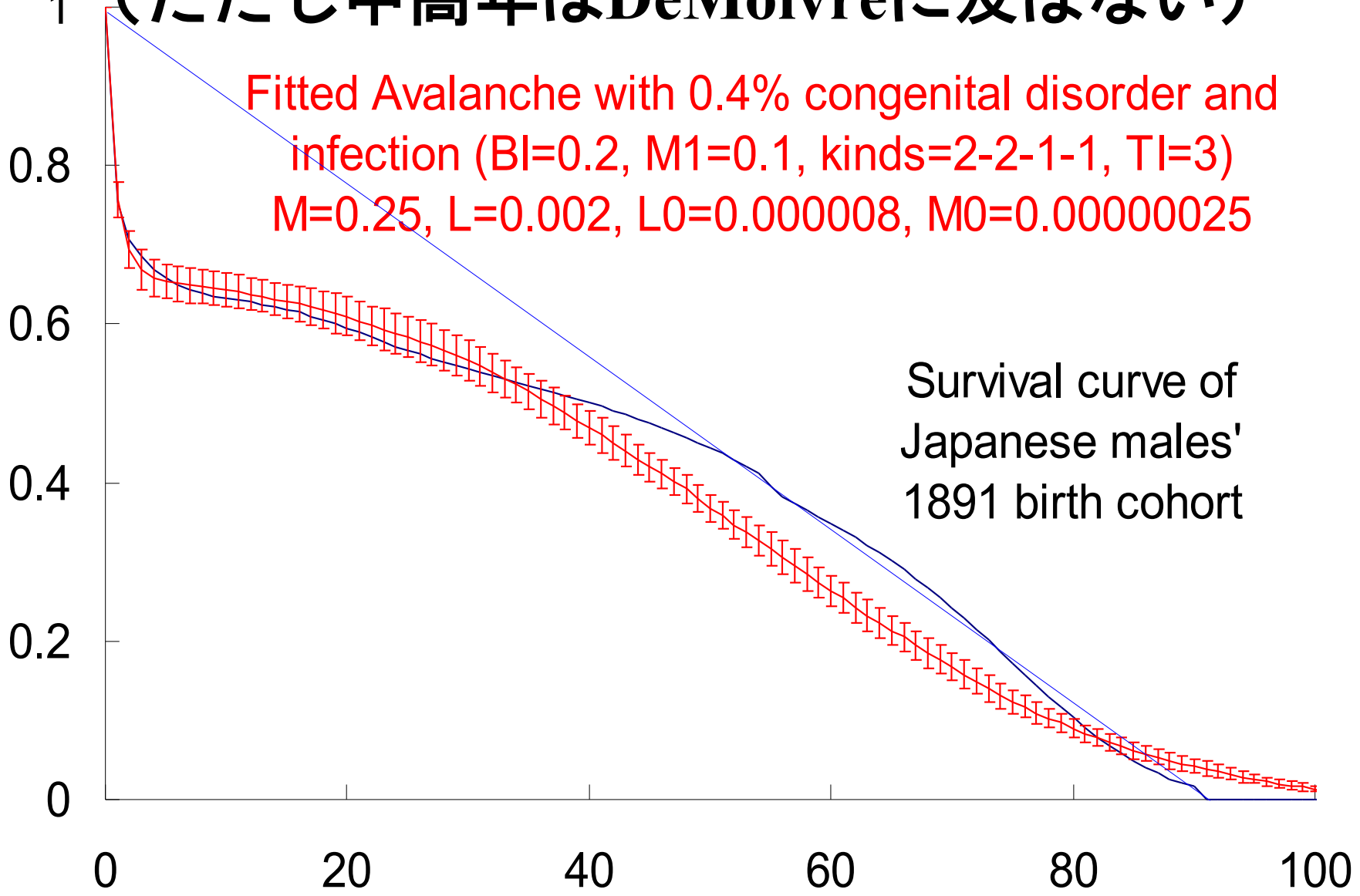
先天異常と感染症を組み込んだ「雪崩」モデル



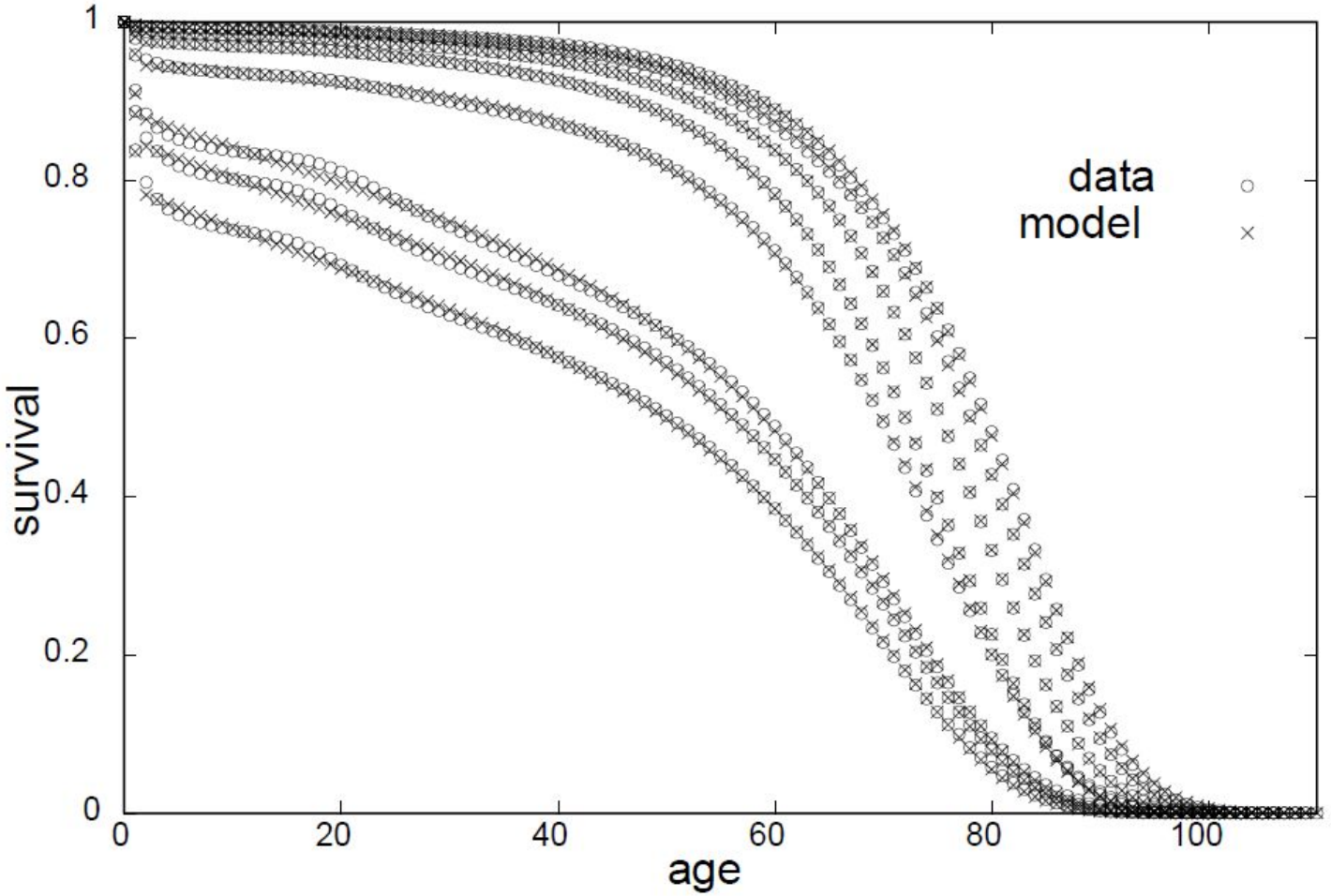
感染症の分類

	致命率が高い	致命率が低い
免疫原性が高い	(例)麻疹	(例)風疹
免疫原性が低い	(例)マラリア	(例)通常のインフルエンザ

感染症を組み込むと，よりよく適合する（ただし中高年はDeMoivreに及ばない）



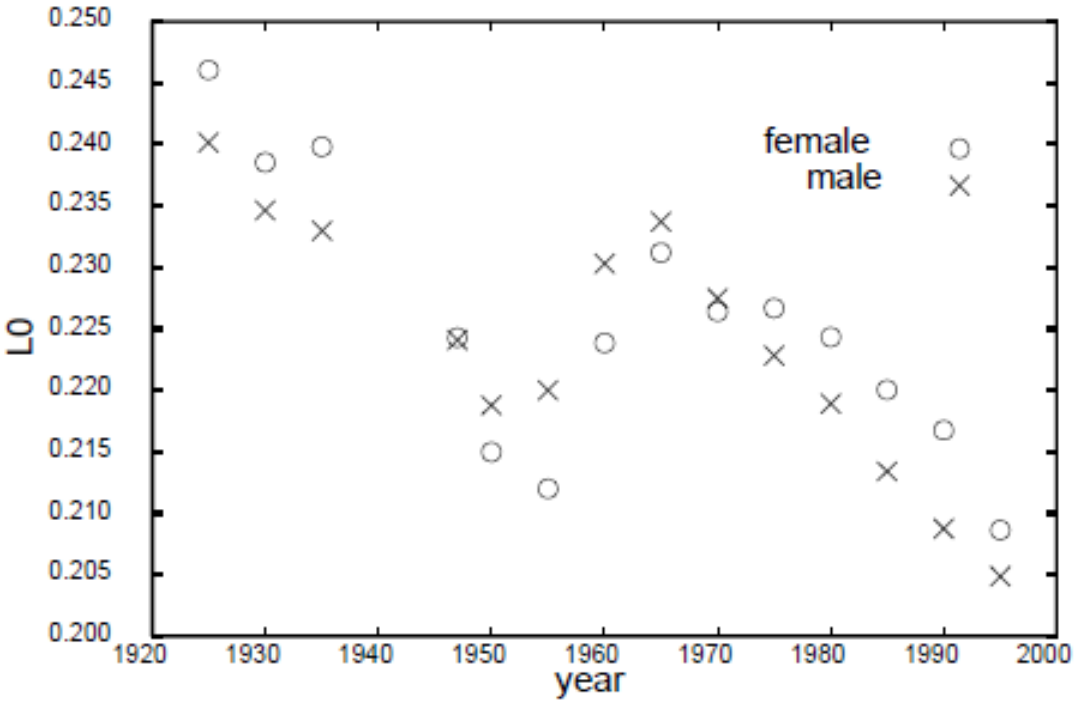
故障と死亡の関係を変える:
dM→{exp(Md)-1}/{exp(MT)-1}



雪崩モデルの利点と欠点

- 利点
  - パラメータに現実的な意味がある。
    - L0は慢性的に人体にダメージを与える外因を意味する（例えば公害など）。栄養状態が良くなれば低下するだろうし脂肪摂取過多なら上昇するだろう
    - Mは慢性疾患の治療水準が上がるにつれて低下するだろう
    - M0は交通事故などによるランダムな死亡水準を意味する
  - 経年的なパラメータの変化を調べると現実の死亡原因の変化が議論できる（次のスライド参照）
  - 個人ベースのシミュレーションに使える
- 欠点
  - M0が変化するとか人口移動があるといった現実に対応できない

パラメータの変化(一部)



- 例えば，このL0の変化は，脳血管疾患による死亡が，戦後から高度経済成長期の，塩分の高い食事，過重労働，過密な居住環境など(それらがL0)によって上昇し，その後，それらの改善によって低下したことと呼応している

２．出生のモデル

出生のミクロモデルとマクロモデル

- ミクロモデル
  - 国立人口問題研究所（当時）の「出生力の生物人口学的分析」(1984)のシミュレーションモデルが代表的。個々の有配偶女性について受胎待ち状態で避妊実行確率，避妊効率，性交頻度，無排卵周期率，中絶確率，産後無月経期間などを確率的に扱う。
  - 際限なし
- マクロモデル
  - Coale and TrussellやHadwigerのものが代表的
  - 中澤（2003)を参照。

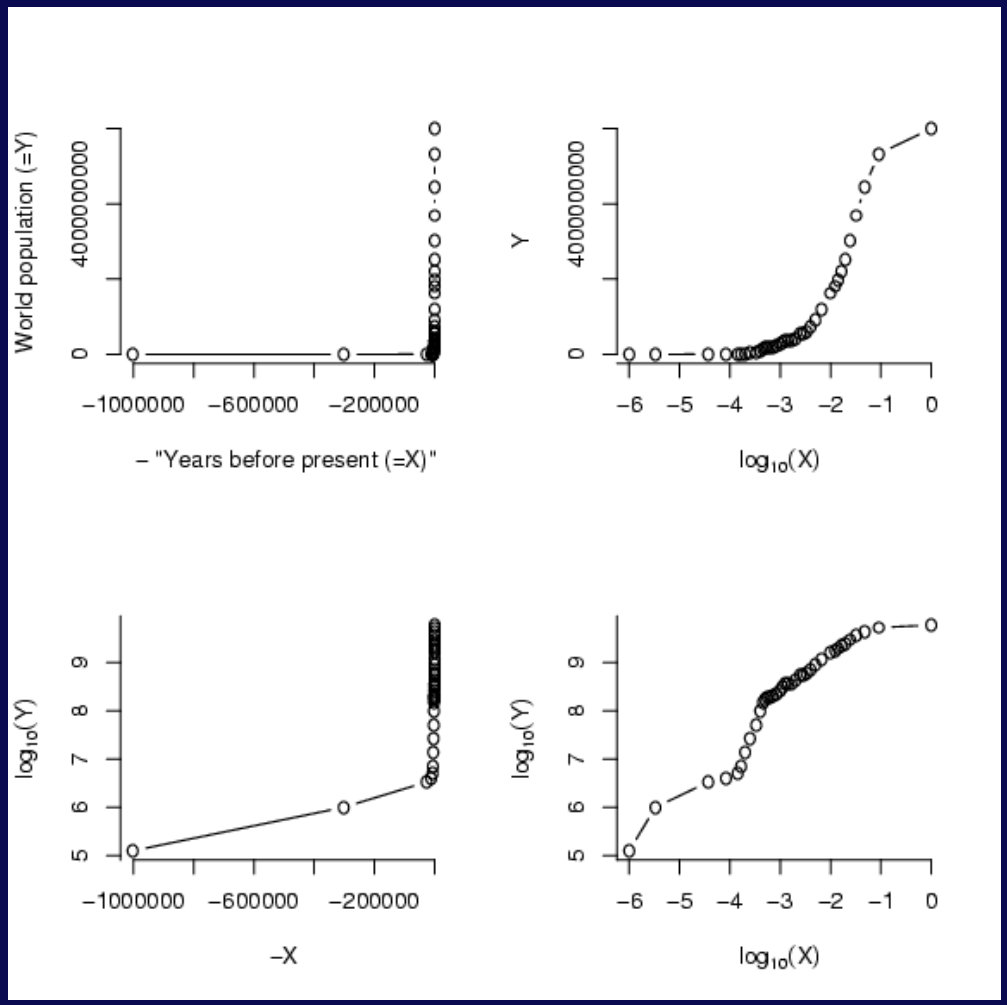
３．人口成長のモデル

世界人口の変化

- 世界人口の変化は時系列データである。
- 時系列で取られているデータは，互いに独立でない。横軸に西暦年をとり，縦軸に世界人口をとってプロットするとき，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ない。
- 時刻を独立変数にした線形回帰は，よく行われるが，一般には正しくない。微分方程式モデルや差分方程式モデルを立てるか，周期性に着目してスペクトル解析を行う方が筋が良い

いろいろなプロット

- 普通に散布図を描いてみると指数関数的に見えるが，片対数や両対数でプロットしてみると，そうではなくて3段階くらいの異なるカーブがつながったものであることがわかる。
- 両対数で線形回帰してその線を延ばして元に戻して予測値を出すのは，よくされているが３重の間違い。
- 微分方程式モデルや差分方程式モデルは原理的には悪くないが，世界人口の変化の場合は，農耕革命や産業革命で微分方程式や差分方程式が変化していると考えられるので限界がある。
- シナリオを仮定してシミュレーションをするのが筋。

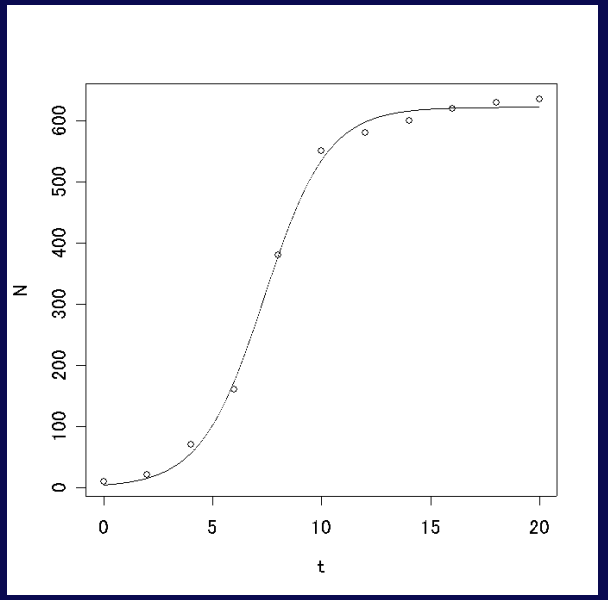


微分方程式/差分方程式モデル

- 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，非線型最小二乗法を用いる。Rではnls()を使うか，optim()で関数を最小化するパラメータを求めるが，適切な初期値を求めるのが難しい
- 世界人口の変化についての微分方程式モデル（時刻をt，人口をN，増加率をr，初期人口をN0，人口収容力をK として）
  - 指数的増加モデル: dN/dt=rN，即ちN=N0 exp(rt)
  - ロジスティック増加モデル: dN/dt=rN(1−N/K)
    即ちN=K/{1+(K/N0−1) exp(−rt)}
  - 最後の審判日モデル: dN/dt=rN^2
  - 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すると考えたもの。先進国の割合をp として，dN/dt=dN1/dt+dp(N−N1)/dt=r1N1+pr2(N−N1)
- 人口は整数なので，微分よりも差分と見るほうが本質的かもしれない。差分方程式ではカオスが起こることもある。
- ただ，人口を数としてだけ見て，中身に踏み込んでいない点が限界

非線形最小二乗法

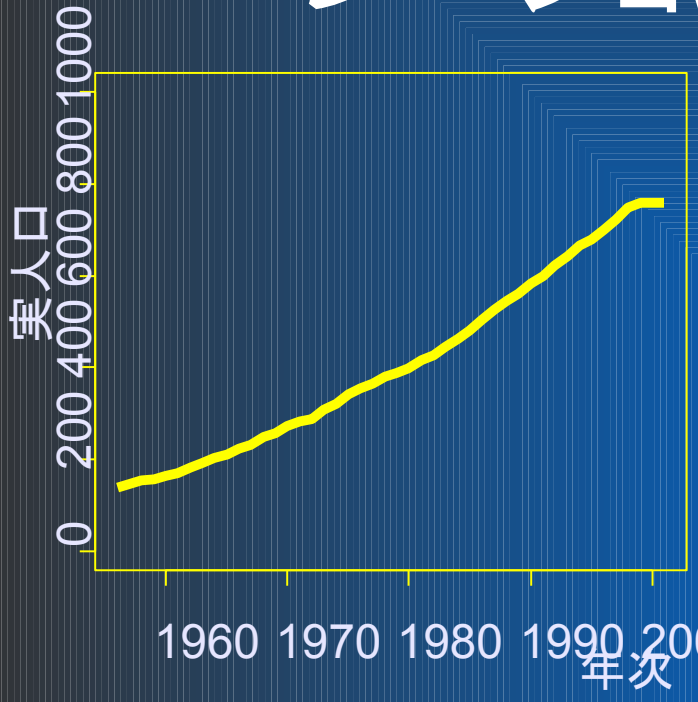
- 世界人口データでは収束しないので，簡単なサンプルデータでやり方を示す。
- 試験管で培養している酵母菌の量の経時的変化のデータが
時間　0　2　4　6　8　10　12　14　16　18　20
量　10　20　70　160　380 550 580　600 620　630　635
であるとき，酵母菌自身が出す有毒物質が環境抵抗となってロジスティック成長していると考えられるので，それを当てはめてみる。
- P <- data.frame(t=seq(0,20,by=2),N=c(10,20,70,160,380,550,580,600,620,630,635))
getInitial(N~SSlogis(t,K,tmid,r),data=P)
res <- nls(N~SSlogis(t,K,tmid,r),data=P)
summary(res)
tt <- seq(0,20,by=0.01)
plot(P)
lines(tt,predict(res,list(t=tt)))

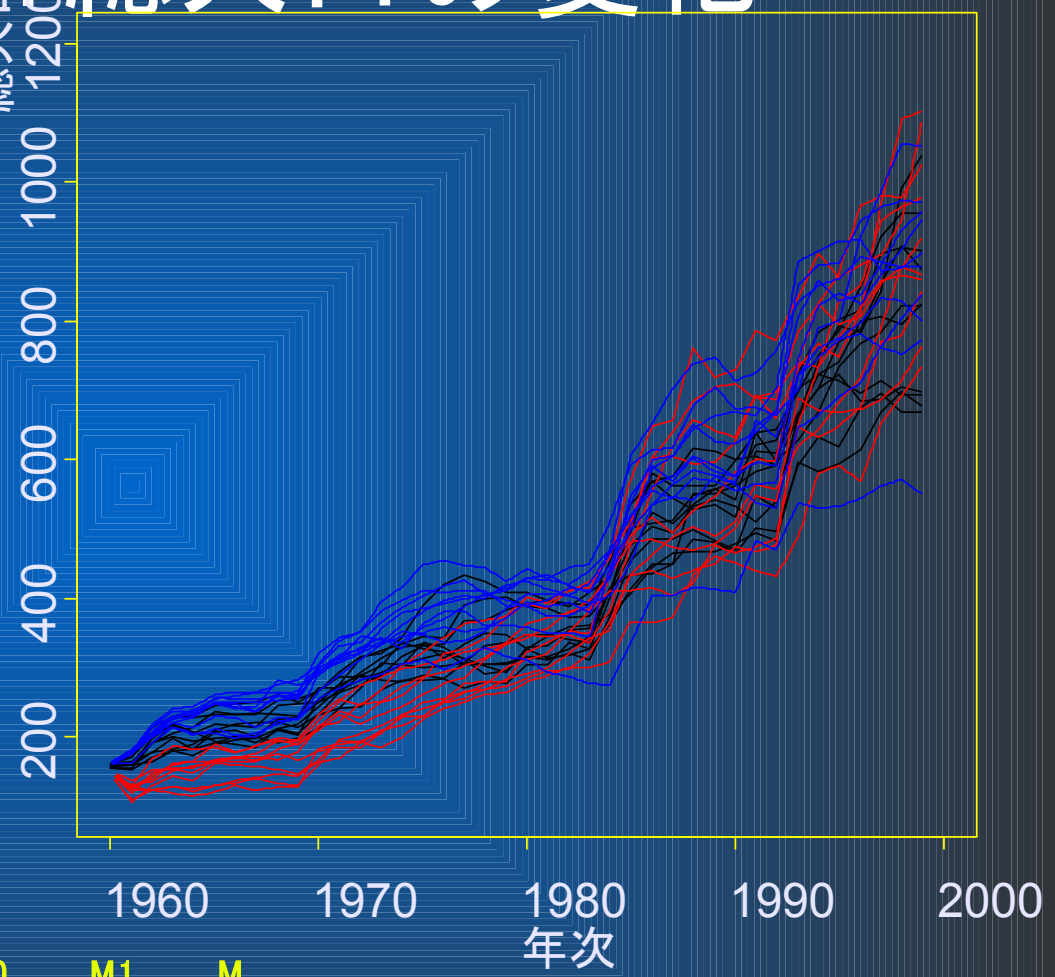


総人口のシミュレーション

- 最低限，出生モデルと死亡モデルを組み合わせる必要がある（cf. 解析的には安定人口モデルが有名）
- 結婚モデルも入れる場合が多い（cf. 解析的には両性安定人口モデルに相当する。数理解析は困難だがシミュレーションなら容易）
- 移動モデルは困難
  - ミクロデータが取りにくい
  - 法則性を仮定しにくい（外的影響を受けやすい）
- マクロな予測にも使われるが，小集団の分析で偶然起こりえたばらつきの幅を推定するのに向いている

ソロモン諸島パラダイス村シミュレーション：総人口の変化





パラメータ	w	L1	L0	L	M0	M1	M
（説明）	初期故障率	初期故障量	偶然故障	failure単位	定率死亡	ランダム死亡	死亡単位
実線(set 1)	0.02	0.2	8.0E-6	0.05	4.0E-6	0.006	0.025
赤破線(set 2)	0.3	0.7	8.0E-6	0.05	2.0E-6	0.002	0.025
青点線(set 3)	0.01	0.1	2.0E-7	0	1.0E-7	0.008	0.004

文献

- Wood JW, Holman DJ, Weiss KM et al. (1992) Hazards model for human population biology. Yearbook of Physical Anthropology, 35: 43-87.
- Mori Y, Nakazawa M (2003) A new simple etiological model of human death. 人口学研究, 33: 27-39.