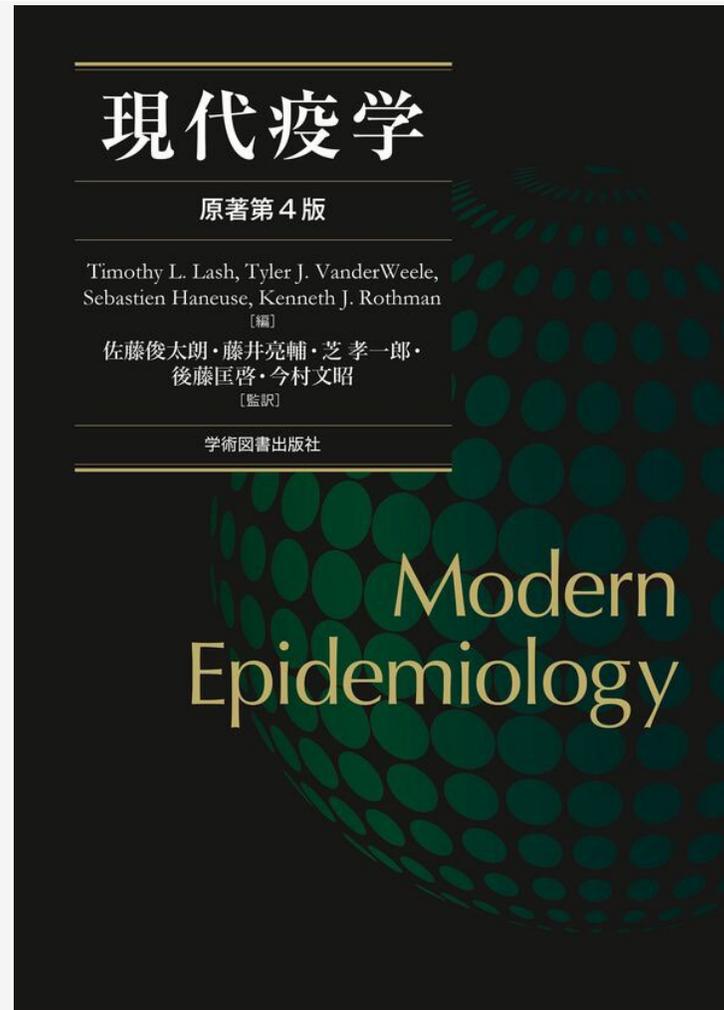


生存時間解析の概要

慶應義塾大学病院 臨床研究推進センター

生物統計部門

長島 健悟 (<https://nshi.jp/>, nshi@keio.jp)



現代疫学 原著第4版. 学術図書出版社 2024.

<https://www.gakujutsu.co.jp/product/978-4-7806-1245-5/>



概要

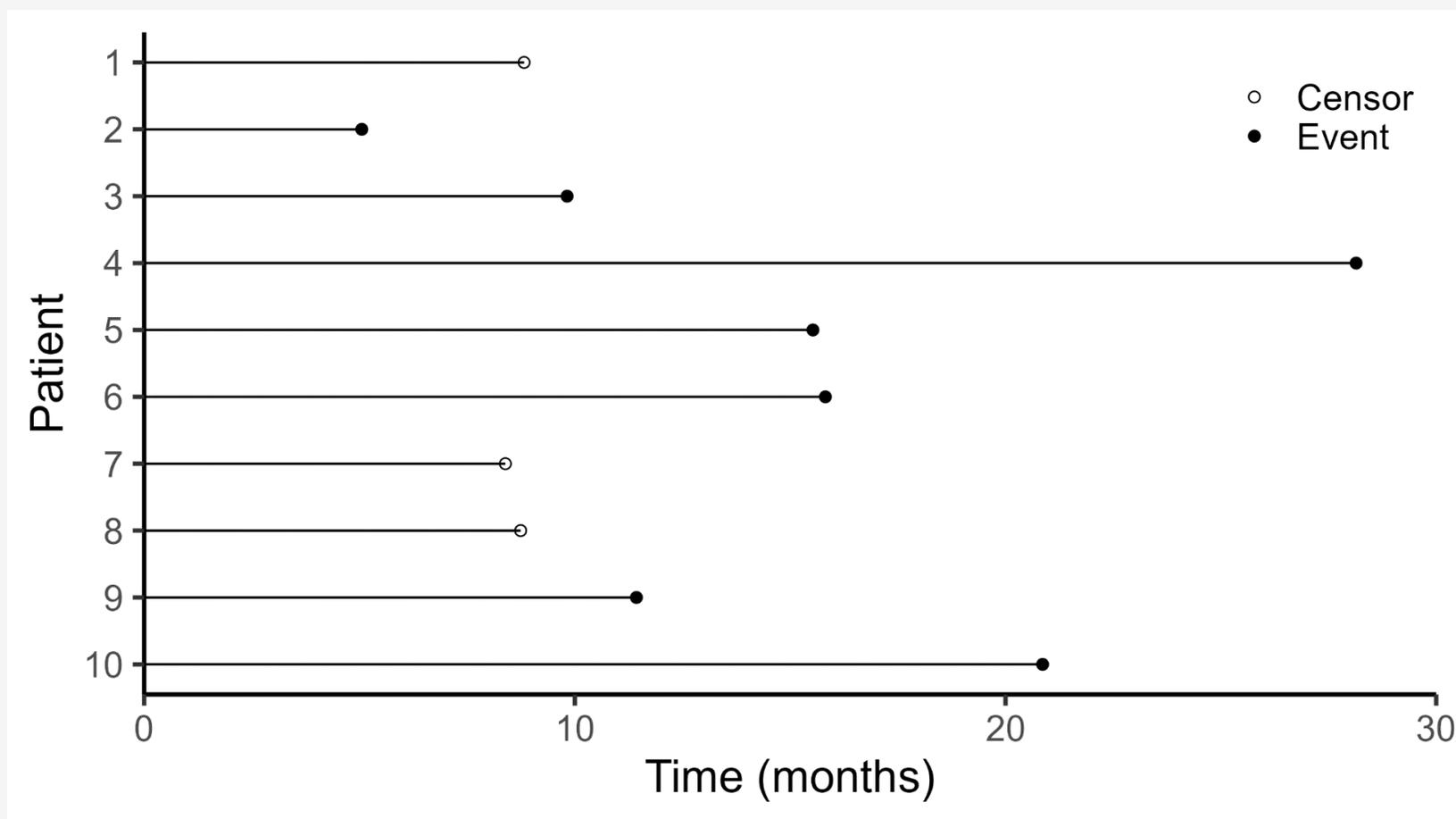
- 生存時間解析の基本概念
- 生存関数の推定
 - Kaplan–Meier法とKaplan–Meierプロット
 - 生存時間中央値と制限付き平均生存時間
- 生存関数の群間比較
 - Logrank検定
- 生存時間アウトカムの回帰モデルほか
 - Cox回帰モデル (Coxの比例ハザードモデル)
 - 加速モデル

生存時間解析

- 生存時間解析
 - 打切りのあるデータの統計的推測理論
 - ある基準時刻から目的の反応（イベント）が起こるまで追跡を行った時間のデータを解析する手法
 - 通常の解析と違い、観測不能例が存在してよい
- 例：がん患者を治療開始時点から再発や死亡までの期間を調べた研究、薬Aと薬Bどちらが再発や死亡を抑えられるか？

生存時間データ

- 起点を0として、起点からイベントまたは打ち切りが観測されるまでの時間のデータ



生存時間データの構造

- 各個人の最終観察時間と打ち切りがあったかどうかの組 (Y_i, δ_i) が観測される
- Y_i : i 番目人の打ち切り時間 C_i とイベント時間 T_i の最小値
 - 実際は両方を観測できないが、全部の人に C_i と T_i があり、小さいほうだけが観測値として得られる
- $\delta_i = 0$: 打ち切り、 $\delta_i = 1$: イベント
 - $C_i = 3, T_i = 10$ の人は $Y_i = 3, \delta_i = 0$
 - $C_i = 12, T_i = 7$ の人は $Y_i = 7, \delta_i = 1$

イベントと打切りと切断

- イベント・打切りのほかに、切断という現象もある
- イベント (event)
 - 興味のある事象：死亡、再発など
- 打切り (censoring)
 - イベントが生じなかった事は分かる
- 切断 (truncation)
 - イベントが生じていない対象を研究から除外

打切り・切断のタイプ

	タイプ	説明
打切り	左側	観察開始前にイベントが発生、いつ発生したのか不明
	右側	観察期間内にイベントが発生しなかったことが分かる
	両側	左側と右側の打切りが両方存在する
	区間	イベントがある時間区間内で発生したことが分かる
切断	左側	イベント発生時間が一定の値を超えた人だけが観察される
	右側	イベントがある時間までに発生した人だけが観察される

(右側) 打切り

- 研究計画で定めた時点での追跡の打切り：運営上の打切り
 - 治療開始から追跡期間を3年とする
 - 定めた期日 (例えば2025年12月31日) で追跡を終了
- その他制御できない打切り
 - 研究参加の途中撤回
 - 退院や来院しなくなってしまい追跡が不能となる

打切りに対する仮定

- 打切りは...
 - イベントの発生を妨げるものではなく、イベントを観測する可能性を妨げるもの ⇒ 打切りとイベント間の独立性の仮定
- イベントの発生が妨げられる場合は、別の方法が必要(競合リスクの解析等)

生存時間解析のモチベーション

- 打ち切りを無視した解析ではバイアスが生じる：打ち切りを除外したり、打ち切り時点を死亡と扱うと×
- 生存時間が長い個体ほど打ち切りを受けやすい傾向
- 生存時間分布の扱い
 - 生存時間は必ず正：右に裾を引いた分布
 - 左右対称な分布を利用した方法との相性が悪め：正規分布を仮定した検定などが使えない

打切りの影響の確認1

- 白血病の寛解維持療法のRCTのデータ (+は打切り)

	再発時間 (週)
プラセボ群	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
試験薬群 (6-MP)	6, 6, 6, 6+, 7, 9+, 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

- プラセボ群の平均再発時間： $\hat{\mu}_0 = \frac{\sum_{i=1}^{21} Y_i}{21} = 8.7$

- 試験薬群の平均再発時間： $\hat{\mu}_1 = \frac{\sum_{i=1}^{21} Y_i}{21} = 17.1$

再発時間を
過少評価

- しかし打切りの人の**実際の再発時間**はもっと長いはず
- 6+：打切られずに観測できた場合の再発時間は6週以上

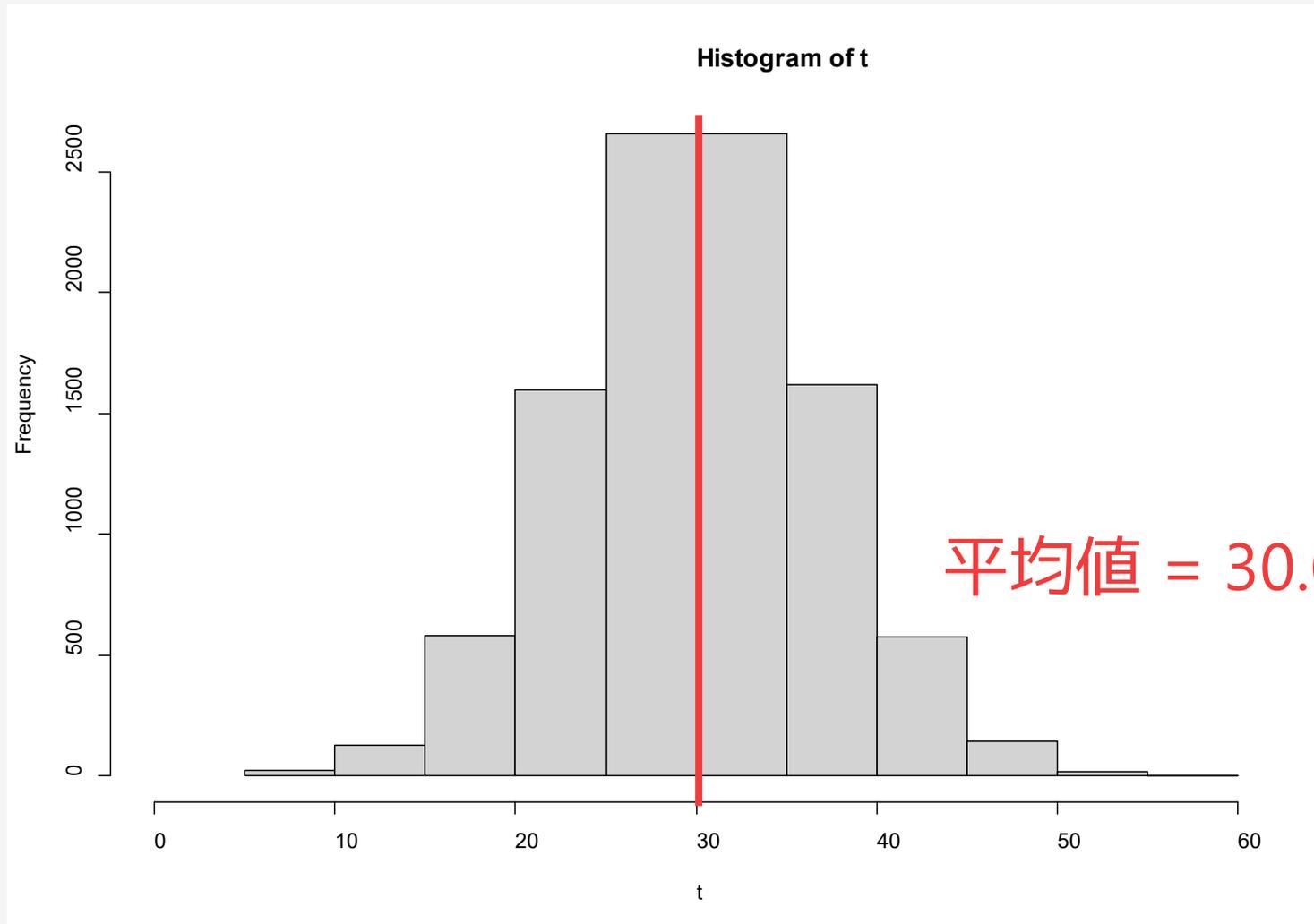
打切りの影響の確認2

- 打切りがないデータと打切りがあるデータで確認
 1. 平均時間が30ヵ月になる打切りがないデータを作成
 2. このデータに対して、人工的に打切りを発生させたデータを作成

二つのデータの分布を要約統計量とヒストグラムを用いて比べてみる

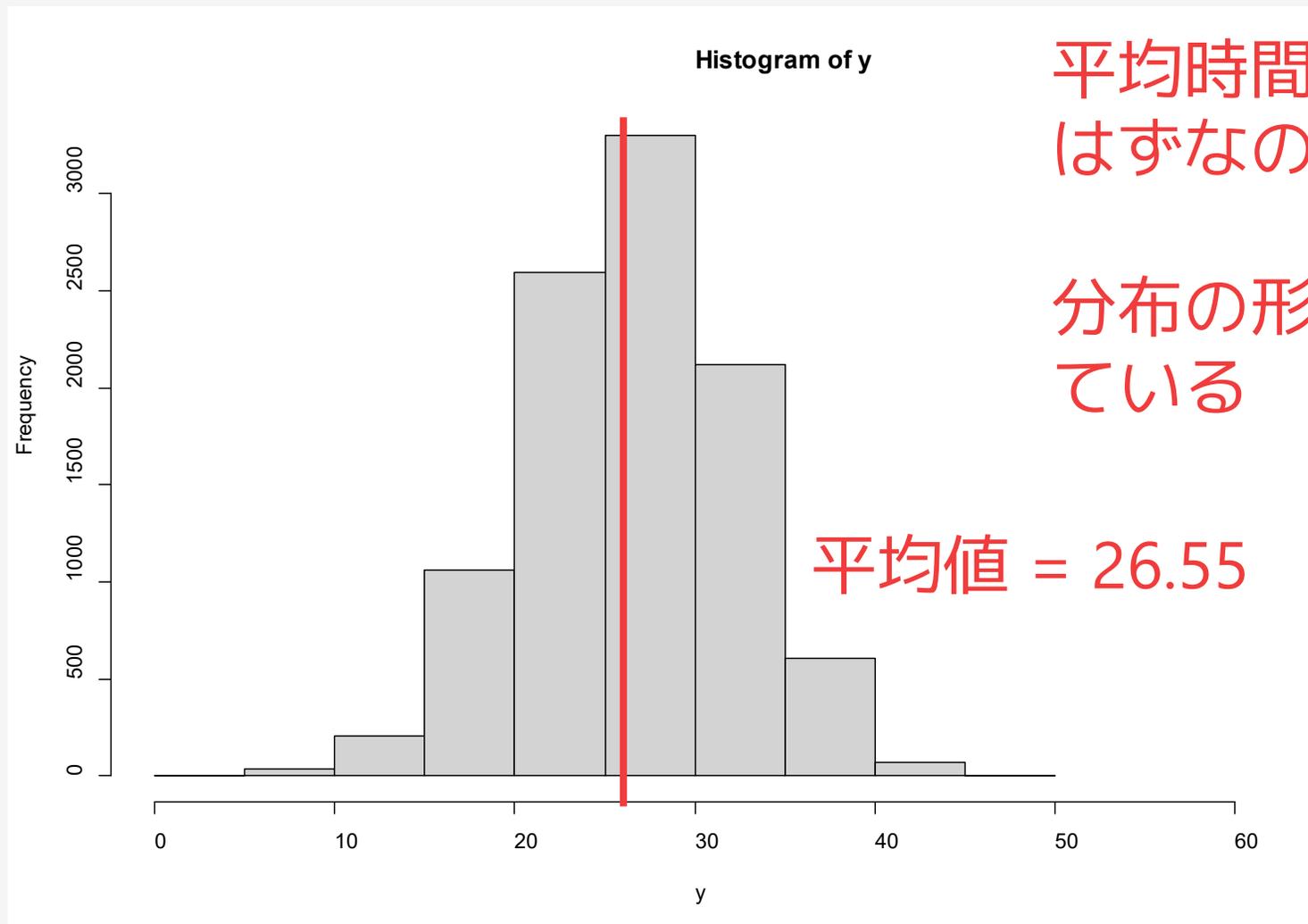
打切りのないデータの結果

- 打切りがないデータ10000個



打切りありのデータの結果

- 打切りがあるデータ10000個



平均時間は30になる
はずなのに・・・

分布の形状も変わっ
ている

打切りデータにおける通常の推測の問題

- うまくいかない：推定結果にバイアス
- 単純な平均値は打切りの情報を考慮していない
 - 標本平均 ×
 - t 検定 ×
 - ノンパラメトリック検定も ×

打切りを扱うために新たな道具が必要

打切りの対処に使う2つの関数

- 生存関数

$$S(t) = \Pr(T > t)$$

- 時点 t を超えるまでイベントが発生しない確率

- ハザード関数

$$\lambda(t) = -\frac{d}{dt} \log S(t)$$

$$\lambda(t) = \lim_{\Delta} \frac{1}{\Delta} \Pr(t \leq T \leq t + \Delta \mid T \geq t)$$

- 時点 t までイベントが発生しなかったという条件のもとでの瞬間イベント発生率

生存関数

- 死亡がイベントなら生存確率
- 再発がイベントなら非再発確率 (recurrence-free)
- 生存関数からAカ月時点の生存確率などが分かる
- 長い期間にわたって生存関数の値が高い (生存確率が高い) 方が平均的な生存時間は長い
- $S_0(t) < S_1(t)$: 1の方が生存状況がよい

ハザード関数

- 死亡がイベントなら死亡率
 - ハザード関数が大きければ死亡しやすい
 - ハザード関数は死亡率 (rate) の指標
- 単純な打切りの場合、生存関数とハザード関数は相互変換可能

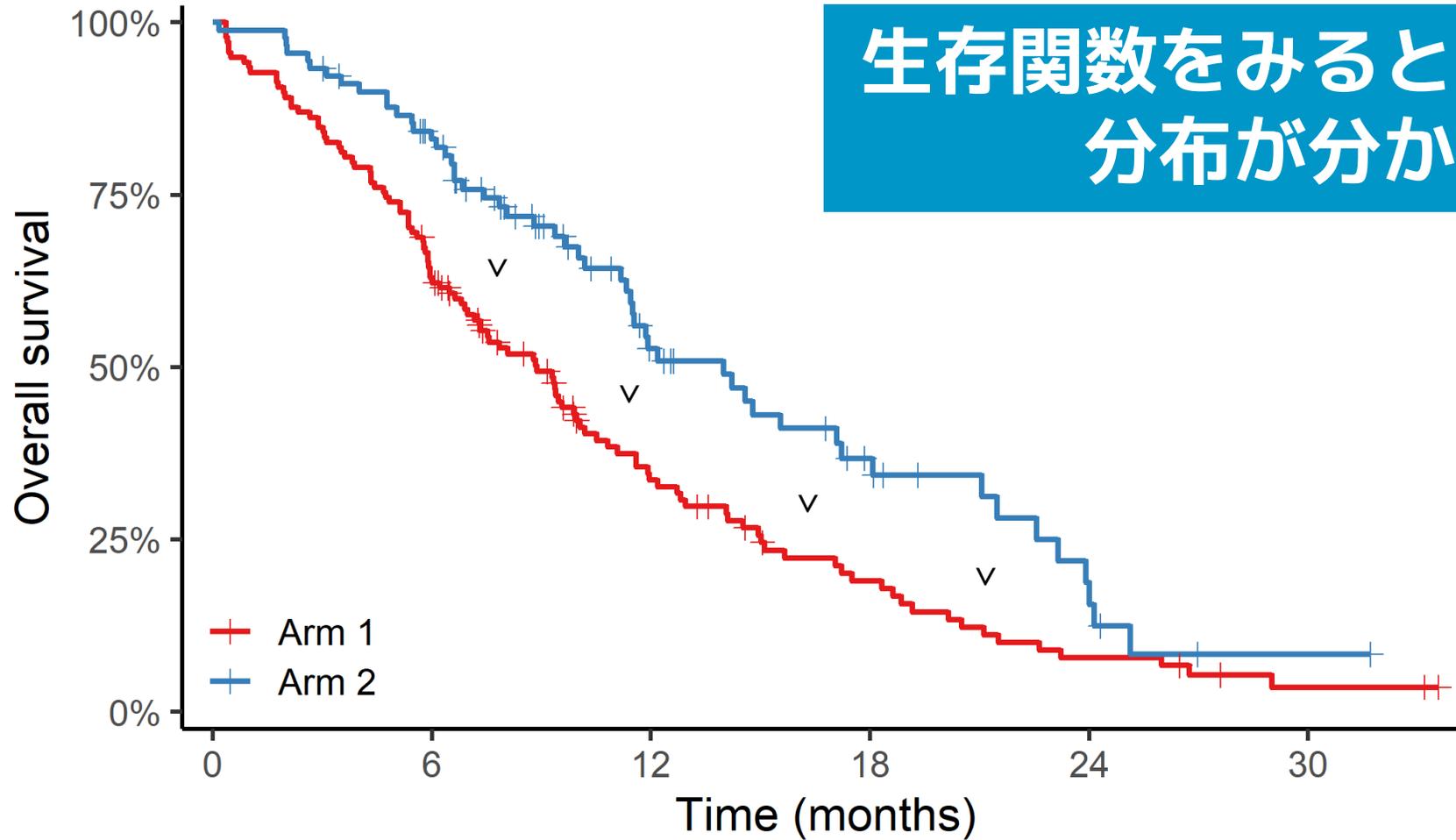
$$\lambda(t) = -\frac{d}{dt} \log S(t)$$

どちらかということ生存関数の方が分かりやすい

ハザード関数については後で登場する

生存関数

生存関数を見るとデータの分布が分かる



	No. at risk	0	6	12	18	24	30
Arm 1	138	86	35	17	7	2	
Arm 2	90	70	30	15	6	1	

概要

- 生存時間解析の基本概念
- 生存関数の推定
 - Kaplan–Meier法とKaplan–Meierプロット
 - 生存時間中央値と制限付き平均生存時間
- 生存関数の群間比較
 - Logrank検定
- 生存時間アウトカムの回帰モデルほか
 - Cox回帰モデル (Coxの比例ハザードモデル)
 - 加速モデル

生存時間データの要約

- 図示
 - **生存関数を推定した結果を直接図にする**
- 要約統計量の算出
 - 生存時間中央値と制限付き平均生存時間
- 白血病の寛解維持療法のRCTのデータを用いる

再発時間 (週)

プラセボ群	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
試験薬群 (6-MP)	6, 6, 6, 6+, 7, 9+, 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

生存関数の推定方法：打切りがない場合

- 時点 t までイベントを生じていない確率なので...

$$\hat{S}(t) = \frac{\{T > t \text{の人数}\}}{N}$$

- プラセボ群の時点 $t = 1, 5, 9$ の生存関数の推定値は？

生存関数の推定方法：打切りがない場合

- 21人中で $t = 1$ までにイベントを生じた人は2人

$$\hat{S}_0(1) = \frac{\{T > t \text{の人数}\}}{N} = \frac{21 - 2}{21} = 90.4\%$$

	再発時間 (週)
プラセボ群	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
試験薬群 (6-MP)	6, 6, 6, 6+, 7, 9+, 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

生存関数の推定方法：打切りがない場合

- 21人中で $t = 5$ までにイベントを生じた人は9人

$$\hat{S}_0(5) = \frac{\{T > t \text{の人数}\}}{N} = \frac{21 - 9}{21} = 57.1\%$$

	再発時間 (週)
プラセボ群	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
試験薬群 (6-MP)	6, 6, 6, 6+, 7, 9+, 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

生存関数の推定方法：打切りがない場合

- 21人中で $t = 9$ までにイベントを生じた人は13人

$$\hat{S}_0(9) = \frac{\{T > t \text{の人数}\}}{N} = \frac{21 - 13}{21} = 38.1\%$$

	再発時間 (週)
プラセボ群	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
試験薬群 (6-MP)	6, 6, 6, 6+, 7, 9+, 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

生存関数の推定方法：打切りがある場合

- 21人中で $t = 9$ までにイベントまたは打切りを生じた人は6人

$$\hat{S}_1(9) = \frac{\{T > t \text{の人数}\}}{N} = \frac{21 - 6}{21} = 71.4\%?$$

- 打切りの2人もイベントと考える ×

	再発時間 (週)
プラセボ群	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
試験薬群 (6-MP)	6, 6, 6, 6+, 7, 9+ , 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

生存関数の推定方法：打切りがある場合

- Kaplan–Meier法と呼ばれる方法を使って、打切りがある場合でも正しい生存関数の推定が可能
- 打切りがある場合はイベントが起こった正確な時間が分からないが
- ある時点までイベントが生じていなかったかどうかは正確にわかる
- この情報をうまく組み合わせると、生存関数の式が出てくる

生存関数の推定方法：Kaplan–Meier法

- 打切りがあると $S(t)$ はデータから直接計算できない
- t の直前までイベントが生じていなかったという条件のもとでの条件付き生存確率は、データから正確に計算できる

$$S(t | T \geq t) = \frac{\{t\text{の時点ちょうどでの生存人数}\}}{\{t\text{の直前までの生存人数}\}}$$

生存関数の推定方法：Kaplan–Meier法

- イベント発生時点による区間 ($0 = t_0 < \dots < t_K = t$) に分割した結果から間接的に $S(t)$ を推定できる

- $$S(t_k | T \geq t_k) = \frac{\Pr(T > t_k)}{\Pr(T \geq t_k)} = \frac{\Pr(T > t_k)}{\Pr(T > t_{k-1})} = \frac{\Pr(Y > t_k)}{\Pr(Y > t_{k-1})}$$

※ 右側打切りのイベント時点ではこの式が成立

$$\begin{aligned} & S(t_k | T \geq t_k) S(t_{k-1} | T \geq t_{k-1}) \cdots S(t_1 | T \geq t_1) \\ &= \frac{S(t_K)}{\cancel{S(t_{K-1})}} \frac{\cancel{S(t_{K-1})}}{\cancel{S(t_{K-2})}} \cdots \frac{\cancel{S(t_1)}}{S(t_0)} = S(t) \end{aligned}$$

$$\because S(t_K) = S(t), S(t_0) = 1$$

生存関数の推定方法：Kaplan–Meier法

- 時点0を t_0 とし各イベントが発生した時点 t_k とする

$$\hat{S}_{KM}(t) = \prod_{k=1}^{L_t} \frac{s_k}{n_k}$$

- L_t : 時点 t までの合計イベント数
- s_k : イベント発生時点 t_k での生存数
- n_k : イベント発生時点 t_k の直前での生存数

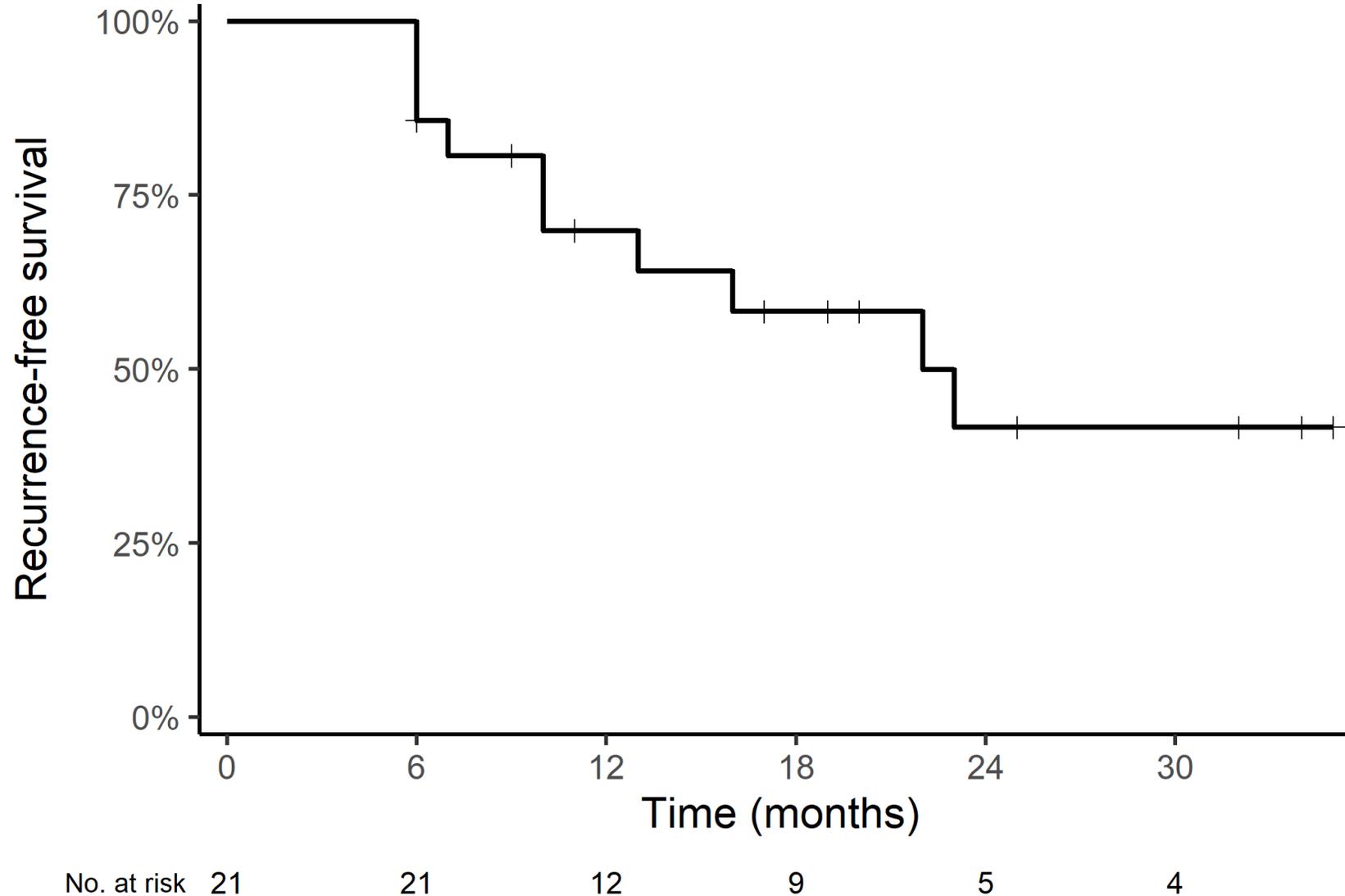
Kaplan–Meier法の計算例

再発時間 (週)

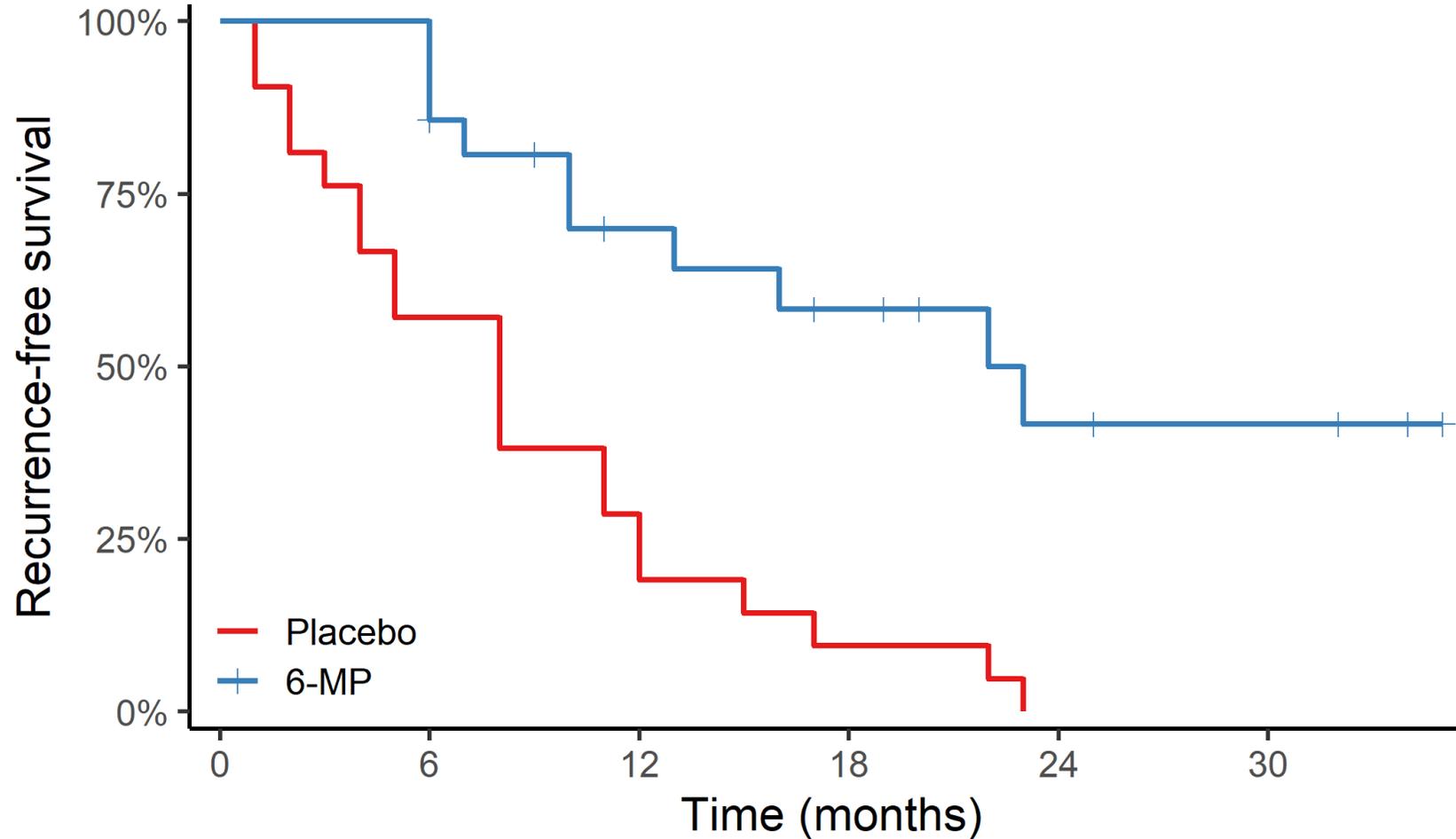
プラセボ群 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
 試験薬群 (6-MP) **6, 6, 6, 6+, 7, 9+**, 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

$t_{(k)}$	n_k	d_k	s_k	s_k/n_k	$\hat{S}_{KM}(t_{(k)})$
0	21	0	21	$21/21 = 1.000$	1.000
6	21	3	18	$18/21 = 0.857$	$1.000 \times 0.857 = 0.857$
7	17	1	16	$16/17 = 0.941$	$0.857 \times 0.941 = 0.806$
10	15	2	13	$13/15 = 0.867$	$0.806 \times 0.867 = 0.699$
13	12	1	11	$11/12 = 0.917$	$0.699 \times 0.917 = 0.641$

Kaplan-Meierプロット



Kaplan-Meierプロット



No. at risk	
Placebo	21 12 6 2 0 0
6-MP	21 21 12 9 5 4

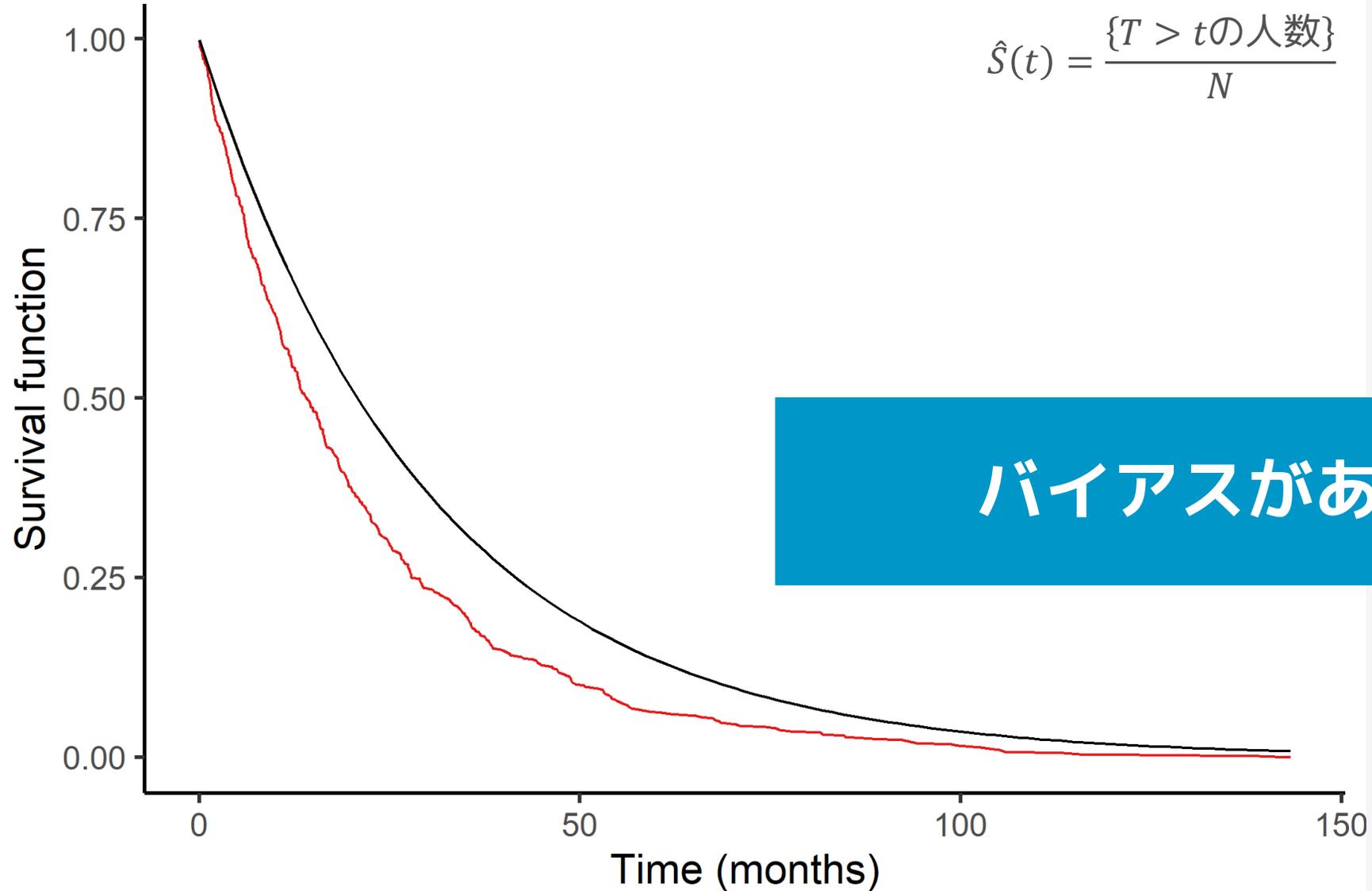
打切りの影響の確認

- 正しい生存関数分かっている打切りがあるデータで

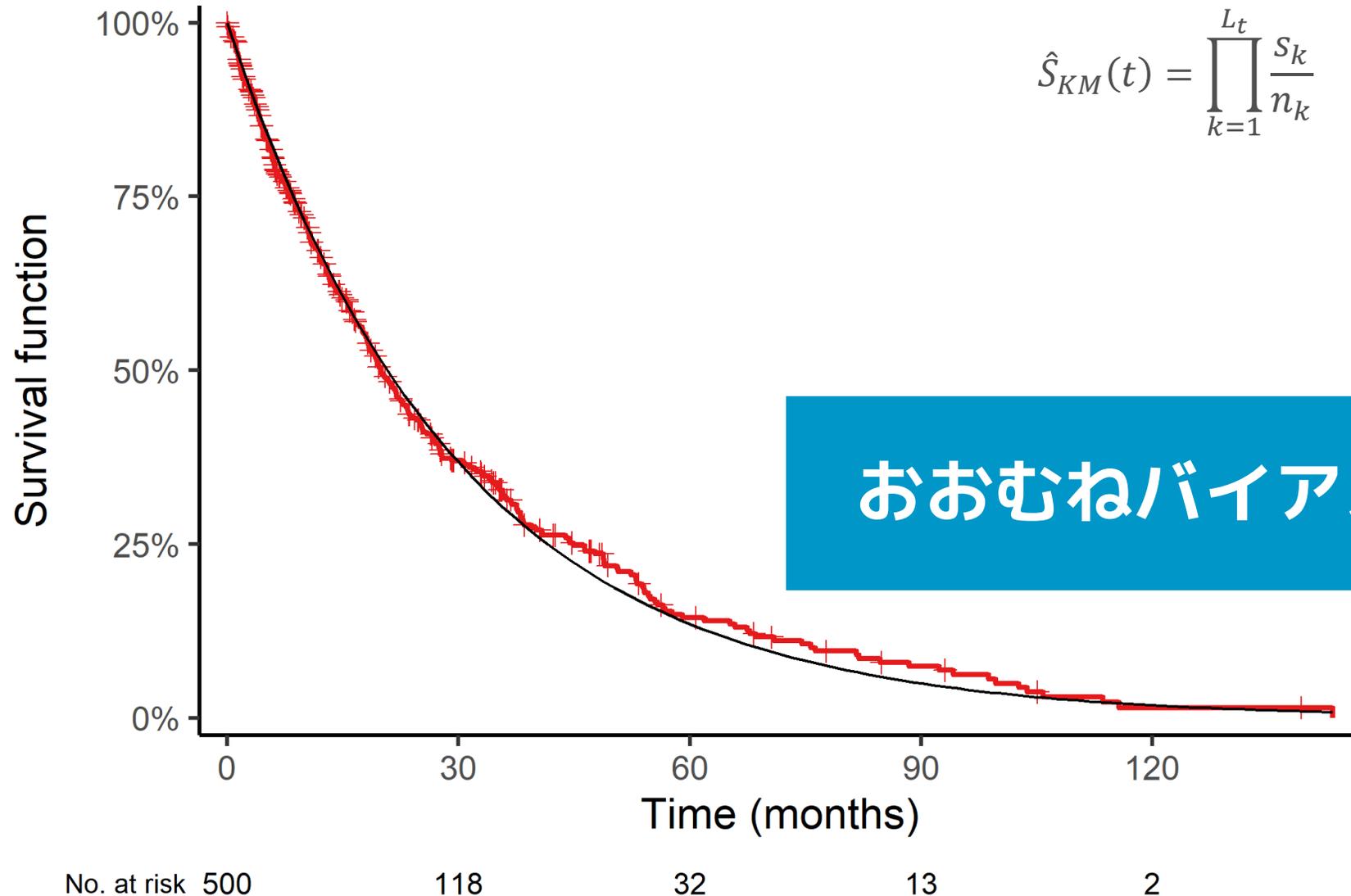
1. $\hat{S}(t) = \frac{\{T>t\text{の人数}\}}{N}$ で推定した生存関数

2. Kaplan–Meier推定量で推定した生存関数
をしてみる

人工データによる確認：打切りを無視



人工データによる確認：Kaplan-Meier法



生存時間データの要約

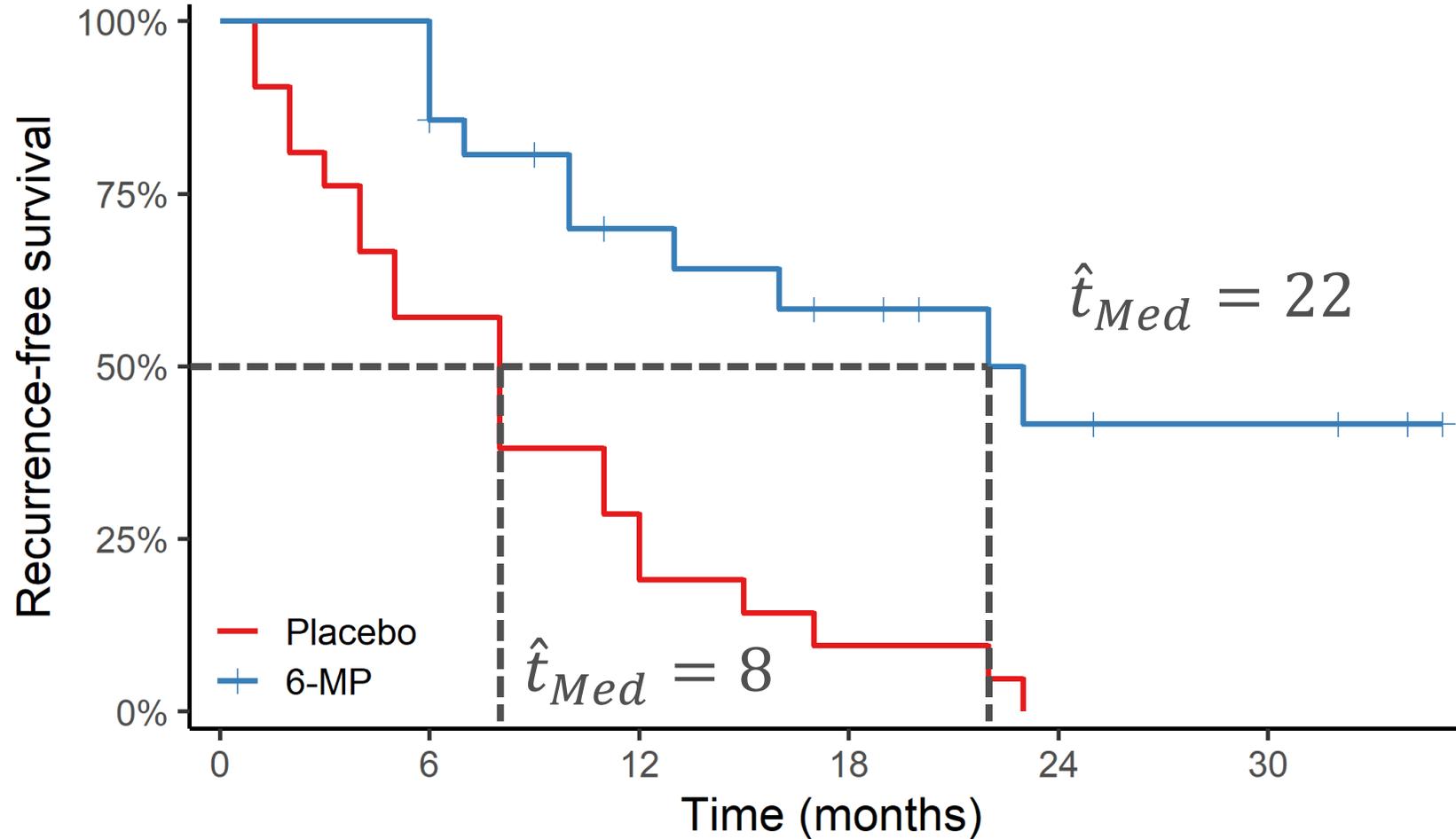
- 図示
 - 生存関数を推定した結果を直接図にする
- 要約統計量の算出
 - **生存時間中央値と制限付き平均生存時間**
- 白血病の寛解維持療法のRCTのデータを用いる

	再発時間 (週)
プラセボ群	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
試験薬群 (6-MP)	6, 6, 6, 6+, 7, 9+, 10, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

生存時間中央値

- 連続な確率変数の場合
 - 平均値で要約する
- 二値の確率変数の場合
 - 割合で要約する
- 生存時間の場合は？
 - 生存時間中央値
 - 集団の50%がイベントを発生する時間
 - $\hat{S}_{KM}(\hat{t}_{Med}) = 0.5$ になる時間 \hat{t}_{Med} が生存時間中央値

生存時間中央值

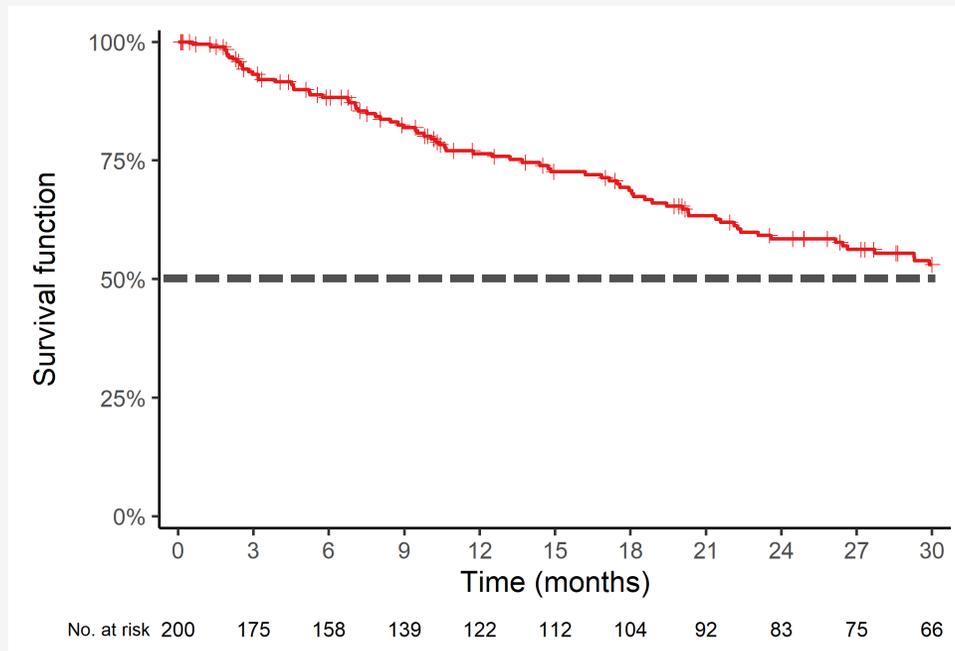


No. at risk

Placebo	21	12	6	2	0	0
6-MP	21	21	12	9	5	4

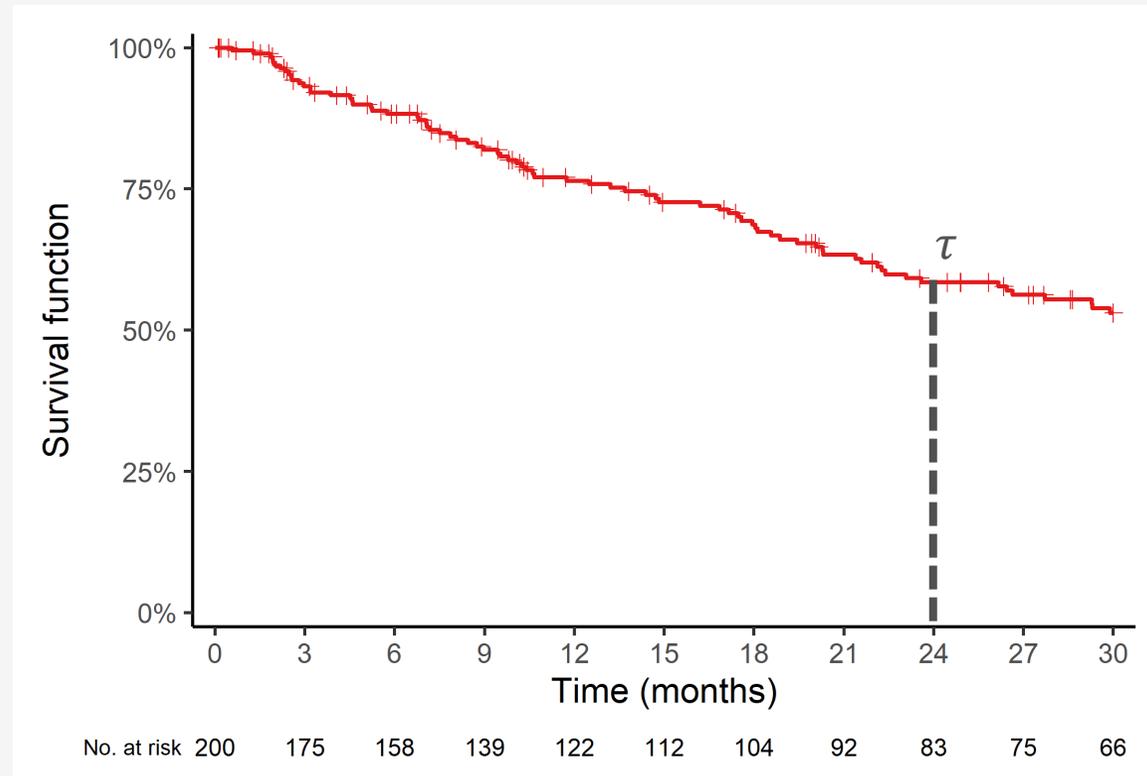
生存時間中央値の問題点

- 最終時点が打切りの場合、Kaplan–Meierプロットは0に到達しない
- イベント数が少ないと集団の50%がイベントを発生しないこともある：生存時間中央値が推定不能



制限付き平均生存時間

- 制限付き平均生存時間 (restricted mean survival time; RMST)
- 定義：ある時点 τ までの生存曲線下面積



制限付き平均生存時間

- Kaplan–Meierプロットの結果から計算できる

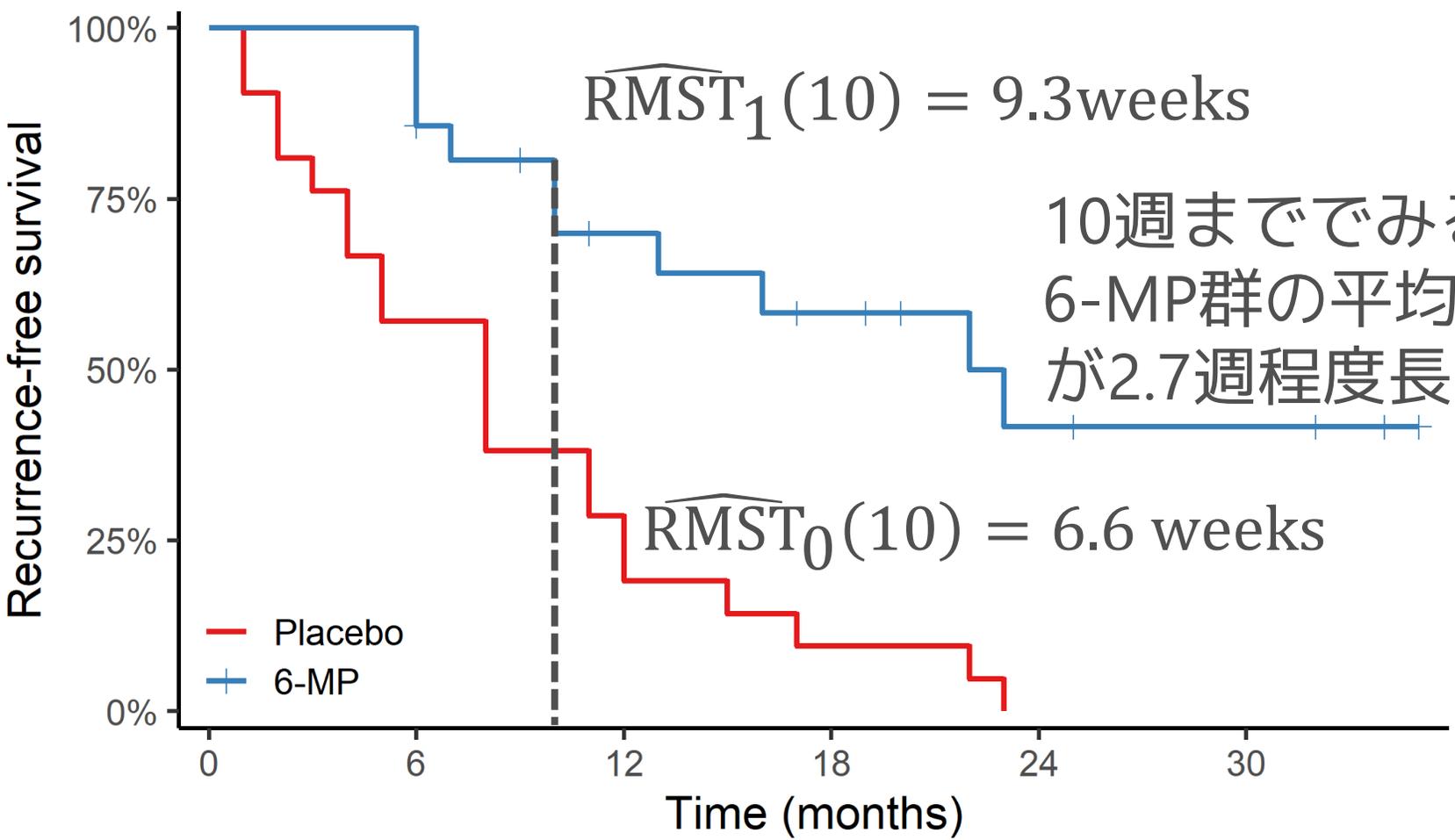
$$\text{RMST}(\tau) = \sum_{k \in t_k < \tau} \hat{S}_{KM}(t_k)(t_k - t_{k-1}) + \hat{S}_{KM}(t_k)(\tau - t_k)$$

RMSTの計算

- $RMST(10) = 0.806 \times (10 - 7) + 0.857 \times (7 - 6) + 1.000 \times (6 - 0) = 9.275$

$t_{(k)}$	n_k	d_k	s_k	s_k/n_k	$\hat{S}_{KM}(t_{(k)})$
0	21	0	21	$21/21 = 1.000$	1.000
6	21	3	18	$18/21 = 0.857$	$1.000 \times 0.857 = 0.857$
7	17	1	16	$16/17 = 0.941$	$0.857 \times 0.941 = 0.806$
10	15	2	13	$13/15 = 0.867$	$0.806 \times 0.867 = 0.699$
13	12	1	11	$11/12 = 0.917$	$0.699 \times 0.917 = 0.641$

RMSTの計算



No. at risk

Placebo	21	12	6	2	0	0
6-MP	21	21	12	9	5	4

$\hat{S}_{KM}(t)$ の標準誤差と信頼区間

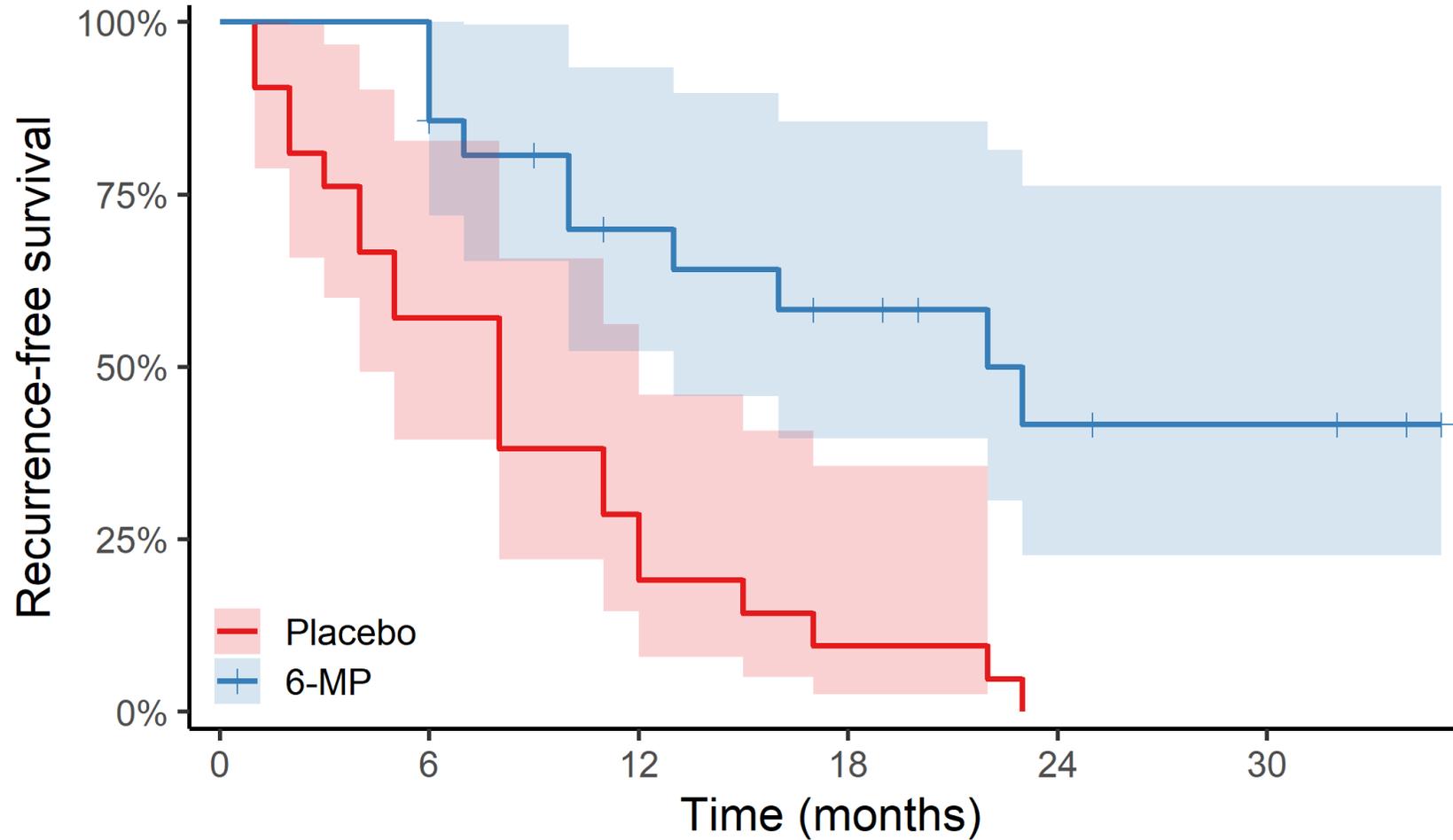
- 標準誤差は以下で求められる

$$\widehat{SE}[\hat{S}_{KM}(t)] = \sqrt{\{\hat{S}_{KM}(t)\}^2 \sum_k^{L_t} \frac{d_k}{n_k s_k}}$$

- 95%信頼区間は以下で求められる

$$\hat{S}_{KM}(t) \pm 1.96 \widehat{SE}[\hat{S}_{KM}(t)]$$

信頼区間を含めたKaplan-Meierプロット



		No. at risk					
		0	6	12	18	24	30
Placebo	21	12	6	2	0	0	
6-MP	21	21	12	9	5	4	

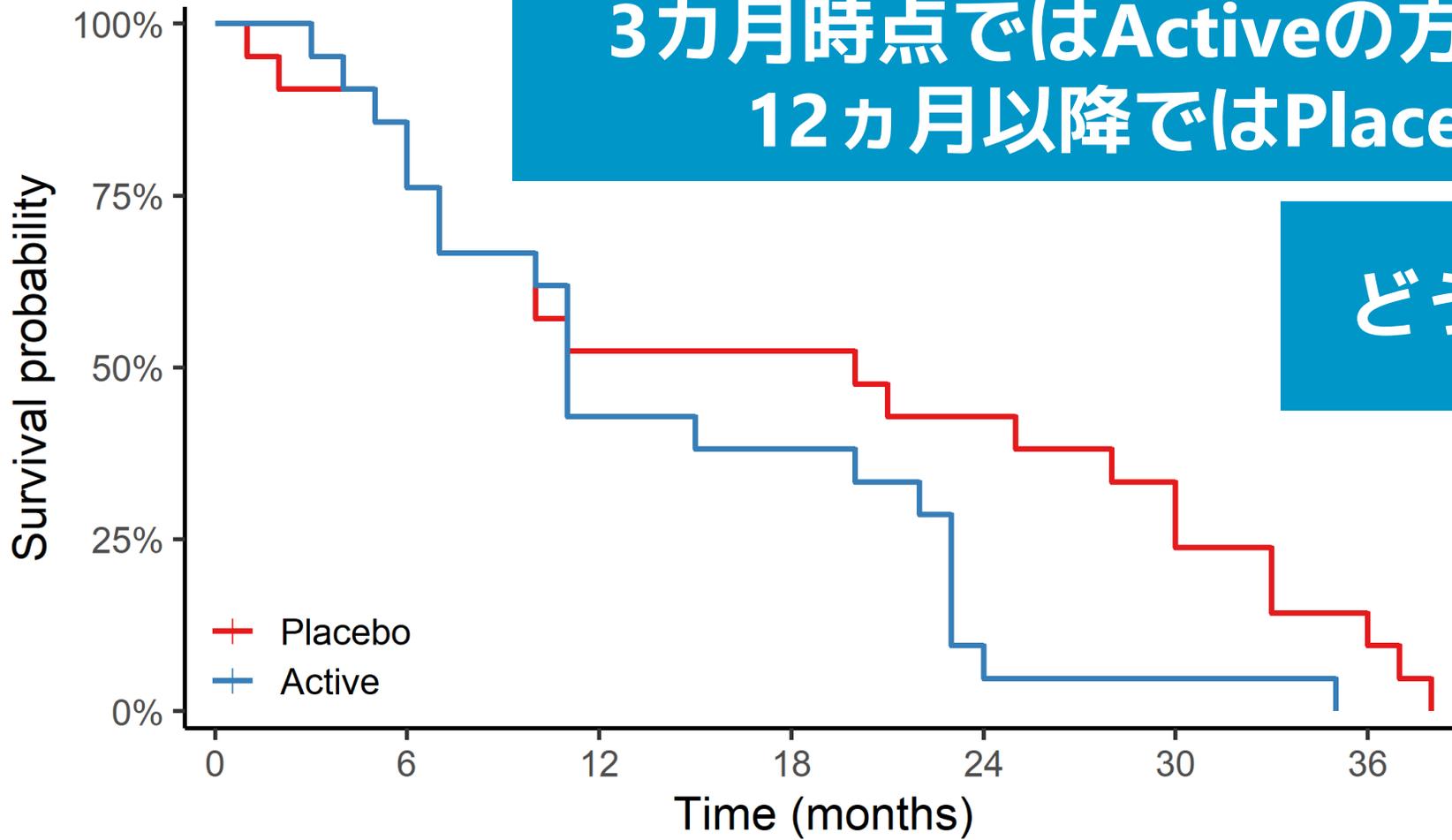
概要

- 生存時間解析の基本概念
- 生存関数の推定
 - Kaplan–Meier法とKaplan–Meierプロット
 - 生存時間中央値と制限付き平均生存時間
- 生存関数の群間比較
 - Logrank検定
- 生存時間アウトカムの回帰モデルほか
 - Cox回帰モデル (Coxの比例ハザードモデル)
 - 加速モデル

生存関数の群間比較

- 臨床的に解釈しやすい時点を定めることができれば、その時点の $\hat{S}_{KM}(t)$ を比較することはできる
 - 群1の5年生存確率 $\hat{S}_{KM,1}(5)$
 - 群2の5年生存確率 $\hat{S}_{KM,2}(5)$
 - 群間差： $\hat{S}_{KM,2}(5) - \hat{S}_{KM,1}(5)$
- 1時点の情報しか使えないためもったいない

生存関数の群間比較



No. at risk

Placebo	21	18	11	11	9	7	3
Active	21	18	9	8	2	1	0

生存関数の群間比較

- Logrank検定 (ログランク検定)
 - 全時点を通じて生存関数に群間差がないかを比較できる仮説検定
- 全時点の情報を使うことができる
- 全時点における差を比べるため、クロスしている場合でも結果は一つになる (統計的な有意差がある or あるとは言えない)

Logrank検定

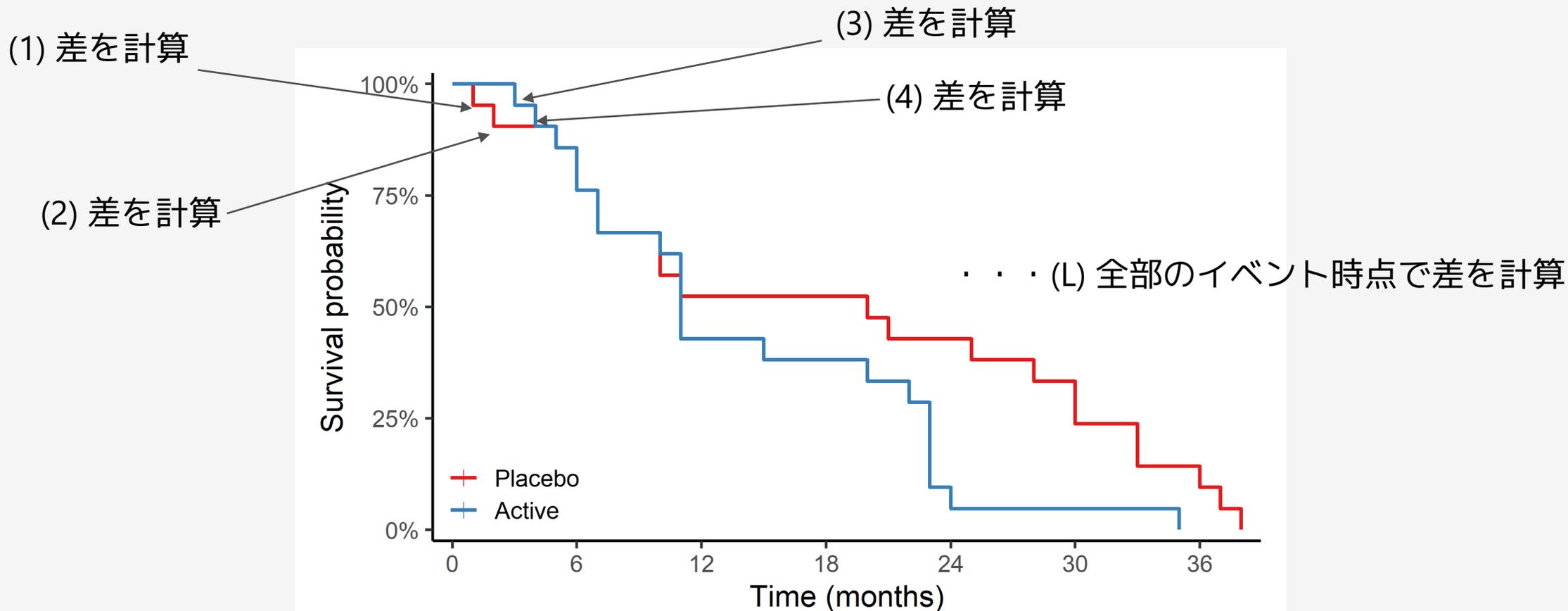
- イベント発生時点の分割表を作って重みをつけて統合

時点 t_k	イベント数	直前までの生存数	合計
群1	d_{1k}	S_{1k}	n_{1k}
群2	d_{2k}	S_{2k}	n_{2k}
全体	d_k	S_k	n_k

- Logrank検定統計量

$$T = \frac{(\sum_{k=1}^L w_k U_k)^2}{\sum_{k=1}^L \text{Var}[w_k U_k]} \sim \chi^2(1), w_k = 1, U_k = d_{1k} - \frac{d_k n_{1k}}{n_k}$$

Logrank検定の直観的な解釈



最後に差の重み付き平均を計算！

9 7 3
2 1 0

Logrank検定

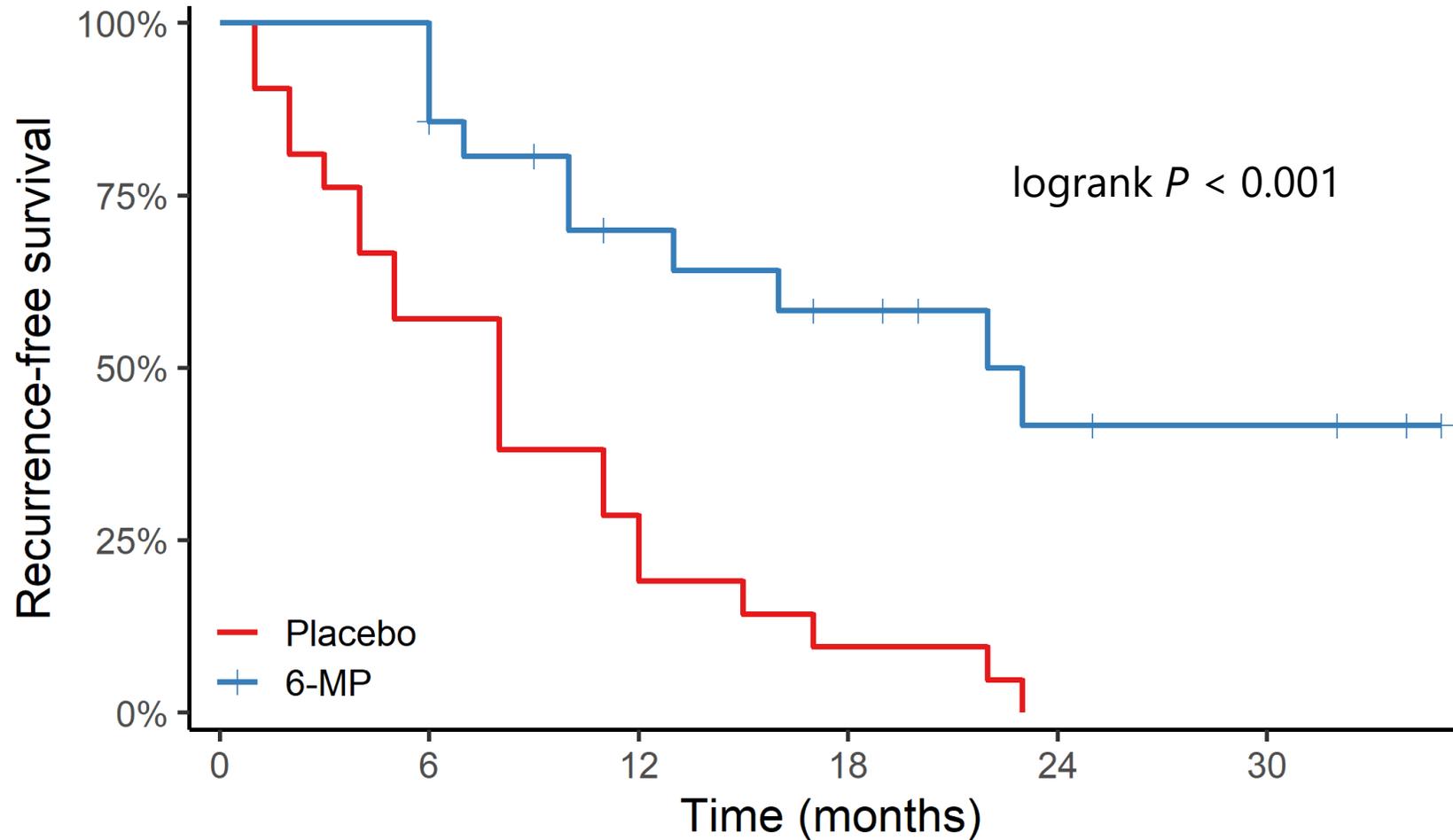
- 分割表の検定と構造は同じ

$$U_k = d_{1k} - \frac{d_k n_{1k}}{n_k} = d_{1k} - e_{1k}$$

- Logrank検定は重みが1になっている
 - 重みを変えた検定もある：Gehanの検定は $w_k = n_k$
- 分散の式はちょっと複雑：ソフトウェアで計算しよう

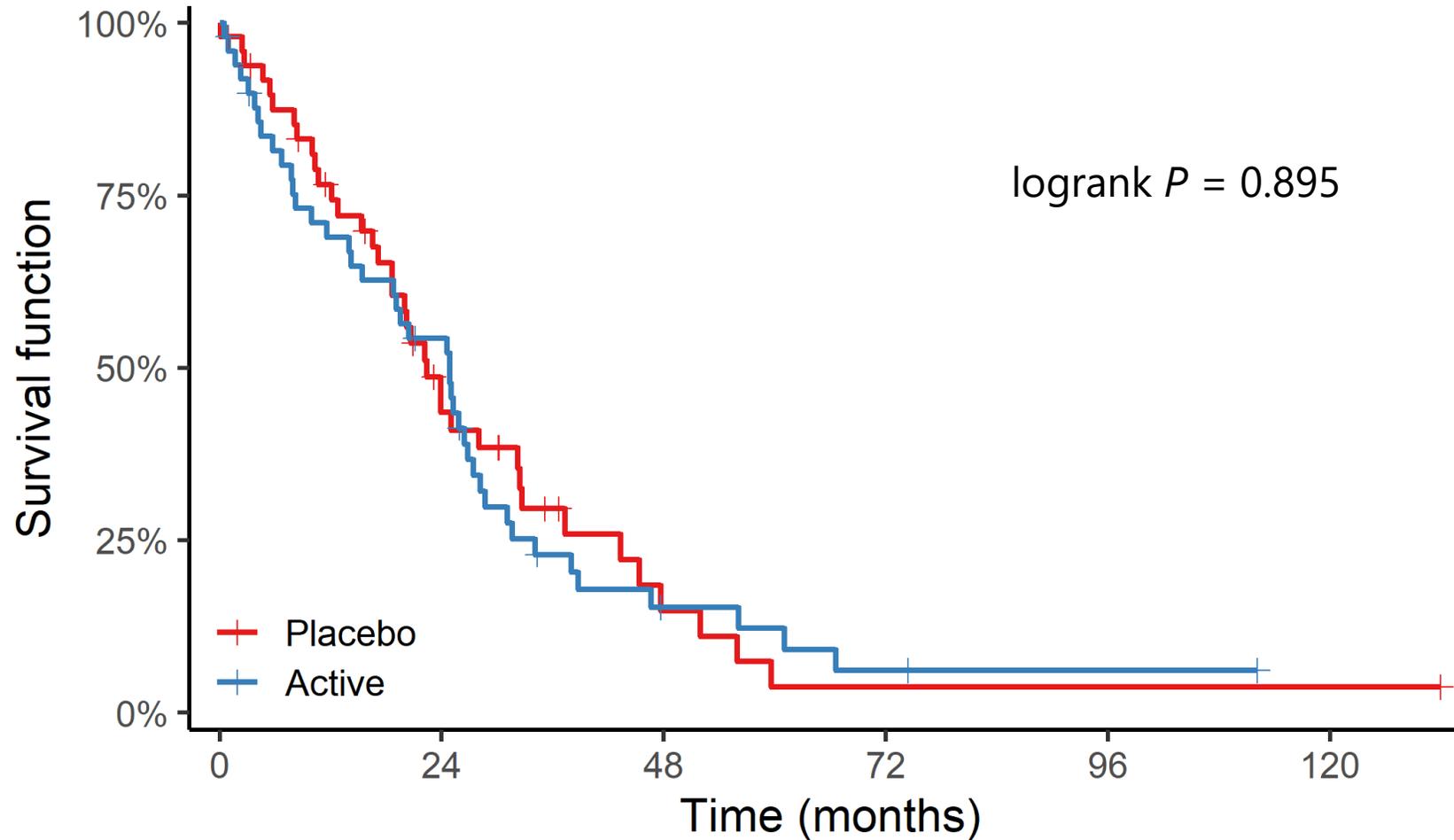
$$\text{Var}[w_k U_k] = w_k^2 \frac{n_{1k} n_{0k} (n_k - d_k) d_k}{n_k^2 (n_k - 1)}$$

Logrank検定の例：白血病RCTのデータ



No. at risk						
Placebo	21	12	6	2	0	0
6-MP	21	21	12	9	5	4

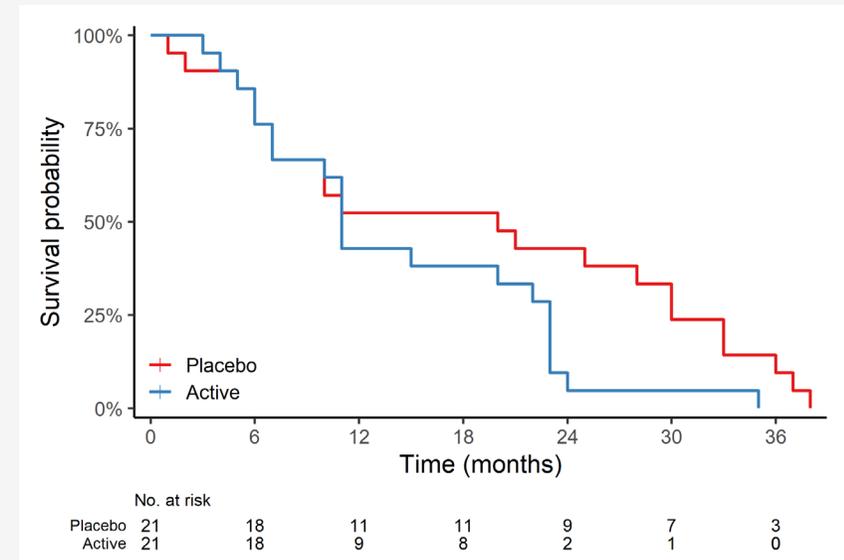
Logrank検定の例：生存曲線が近い場合



		No. at risk					
	0	24	48	72	96	120	
Placebo	50	17	4	1	1	1	
Active	50	25	5	2	1	0	

検定に関する補足

- RMSTもある時点 τ までの生存関数全体を要約した指標
 - 中央値は1時点の情報 $+\alpha$ を利用
- RMSTの差や比の検定は行うことができる
- 特に最初にくっついていた生存関数が後半で差が開く場合にはRMSTの検定の方が望ましい場合もある



ここまでのまとめ

- 生存時間解析の考え方
- 生存関数の推定：Kaplan–Meierプロット
- 要約統計量：生存時間中央値、制限付き平均生存時間
- 群間比較：logrank検定
- 生存時間解析は時間変化するアウトカムを扱うため少し複雑で難しい
- 他のタイプの解析に近い部分もあるため、似ている部分を参考にしながら理解を深めると良い

概要

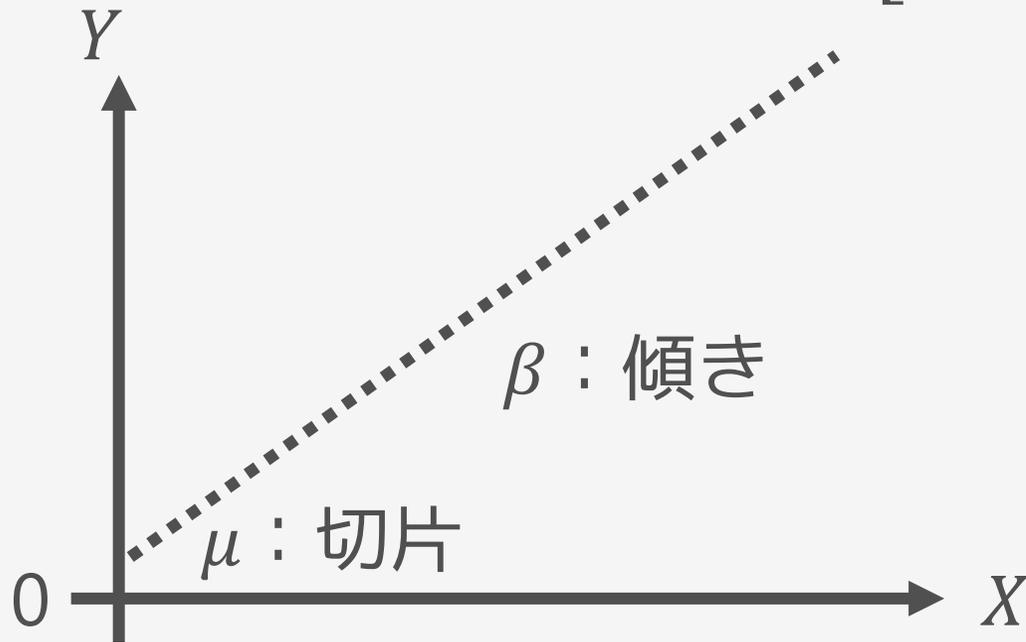
- 生存時間解析の基本概念
- 生存関数の推定
 - Kaplan–Meier法とKaplan–Meierプロット
 - 生存時間中央値と制限付き平均生存時間
- 生存関数の群間比較
 - Logrank検定
- 生存時間アウトカムの回帰モデルほか
 - Cox回帰モデル (Coxの比例ハザードモデル)
 - 加速モデル

回帰モデルの基本

- 結果変数 Y と説明変数 X の間に数式を当てはめたもの
- 線形回帰モデル

$$Y = \mu + \beta \times X + \epsilon, \quad E[\epsilon] = 0$$

$$\Rightarrow E[Y | X] = \mu + \beta \times X$$



- ※ X をある値に固定したときの Y の平均値
これが X の関数で表現できることを仮定
- ※ $E[Y]$ は Y の期待値を表す記号である
- ※ $E[Y | X]$ は X をある値に固定した期待値

**実データには誤差があり
全部が直線上には乗らず**

生存時間アウトカムの回帰モデル？

- 生存時間アウトカムのデータ
- 各個人の最終観察時間と打ち切りがあったかどうかの組 (Y_i, δ_i) が観測される
 - Y_i : i 番目人の打ち切り時間 C_i とイベント時間 T_i の小さい方、 $\delta_i = 0$: 打ち切り、 $\delta_i = 1$: イベント

最終観察時間 Y_i で回帰モデルを作る？
イベント発生の有無 δ_i で回帰モデルを作る？
両方で作る？

打切りの対処に使う2つの関数

- **生存関数**

$$S(t) = \Pr(T > t)$$

- 時点 t を超えるまでイベントが発生しない確率

- **ハザード関数**

$$\lambda(t) = -\frac{d}{dt} \log S(t)$$

$$\lambda(t) = \lim_{\Delta} \frac{1}{\Delta} \Pr(t \leq T \leq t + \Delta \mid T \geq t)$$

- 時点 t までイベントが発生しなかったという条件のもとでの瞬間イベント発生率

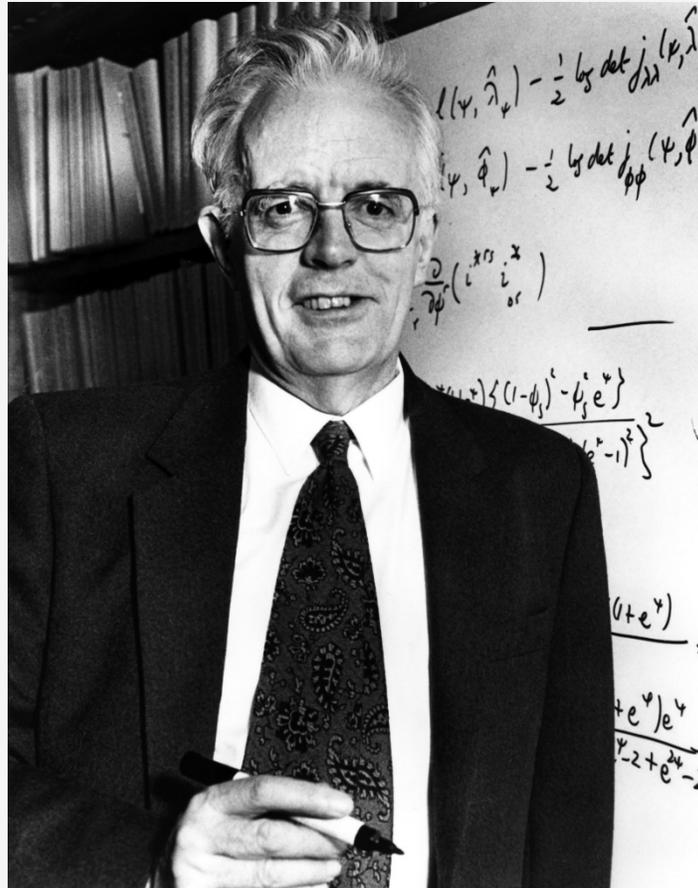
生存時間アウトカムの回帰モデル

- 生存関数の回帰モデル
 - 加速モデル
- **ハザード関数の回帰モデル**
 - **Cox回帰モデル (Coxの比例ハザードモデル)**

どちらの回帰モデルでも使うことができる
※医学領域ではCox回帰モデルが頻用される

Cox回帰モデル (Coxの比例ハザードモデル)

- 1972年にSir David Coxが提案：ハザード関数に対する回帰モデル



Sir David Cox (1924–2022)

Title: Historical Portrait: Sir David Cox
Author: Unknown
Public Domain
<https://commons.wikimedia.org/w/index.php?curid=10570802>

ハザード関数

- 死亡がイベントなら死亡率
 - ハザード関数が大きければ死亡しやすい
 - ハザード関数は死亡率 (rate) の指標
- 生存関数と相互変換可： $\lambda(t) = -\frac{d}{dt} \log S(t)$

他の回帰モデルと同様に
Xを固定したときの平均的な $\lambda(t)$ が異なるような
構造を持つ回帰モデルを考えれば良い

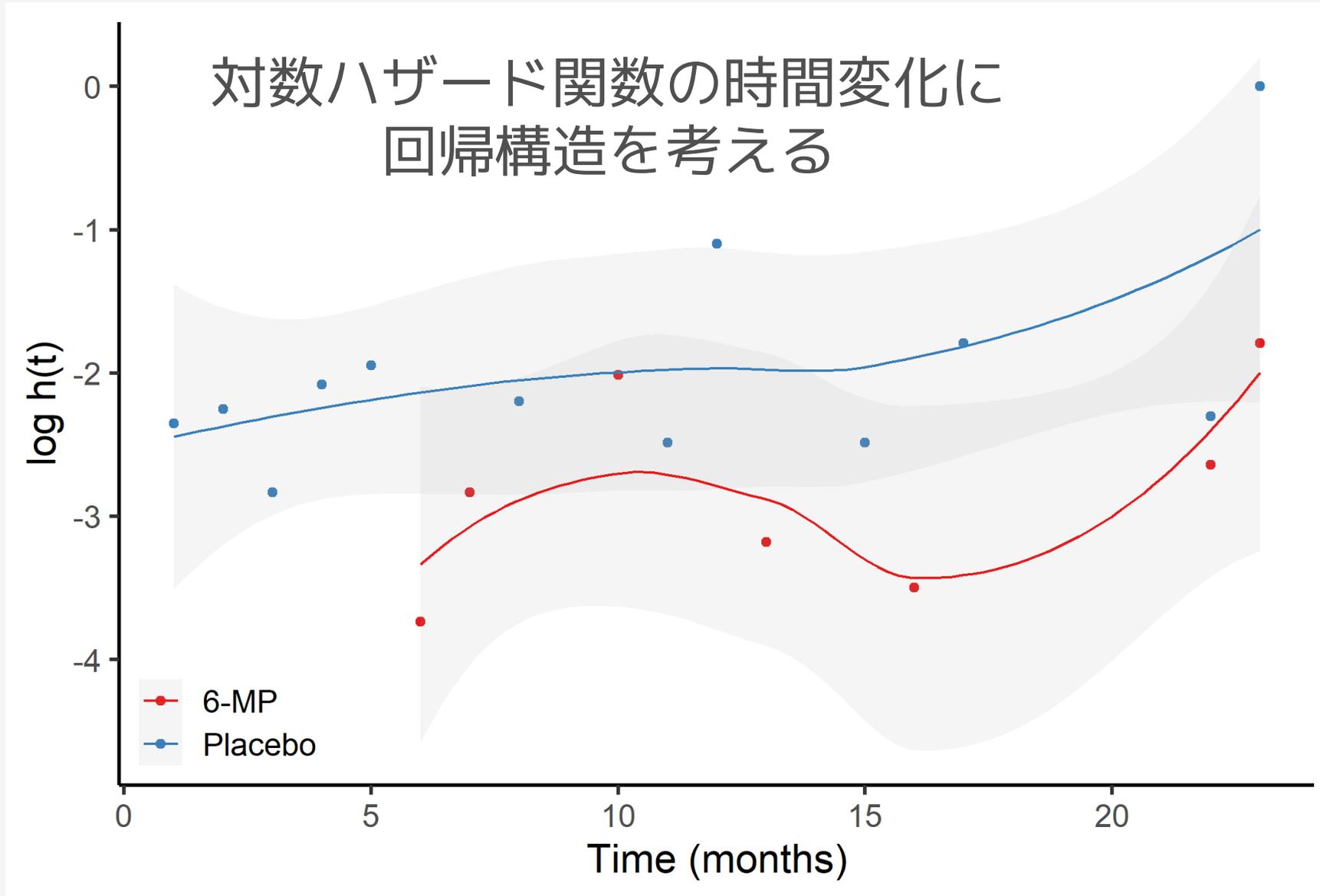
ハザード関数に対する回帰モデル

- ハザード関数に対して以下のモデルを仮定

$$\lambda(t | \mathbf{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

- $\mathbf{X} = (X_1, X_2, \dots, X_p)$: 説明変数
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$: 回帰パラメータ
- $\lambda_0(t)$: 基準ハザード関数

生存関数から推定した対数ハザード関数



基準ハザード関数

- 基準ハザード関数は全ての \mathbf{X} が0の集団におけるハザード関数

$$\begin{aligned}\lambda(t \mid \mathbf{X} = \mathbf{0}) &= \lambda_0(t) \exp(\beta_1 0 + \beta_2 0 + \cdots + \beta_p 0) \\ &= \lambda_0(t) \exp(0) = \lambda_0(t) \times 1 \\ &= \lambda_0(t)\end{aligned}$$

- 基準ハザード関数は線形回帰モデルの切片のような役割を持つ

- $E[Y \mid \mathbf{X}] = \mu + \beta_1 \times X_1 + \cdots + \beta_p \times X_p$

ハザード比とその解釈

- 説明変数が1つ (β_1 のみ) で二値の場合を考える
 - $X_1 = 0$ or $X_1 = 1$
- 二つのハザード関数の比を取ると

$$\begin{aligned} HR &= \frac{\lambda(t \mid X_1 = 1)}{\lambda(t \mid X_1 = 0)} = \frac{\lambda_0(t) \exp(\beta_1 \times 1)}{\lambda_0(t) \exp(\beta_1 \times 0)} \\ &= \frac{\lambda_0(t) \exp(\beta_1)}{\lambda_0(t) \exp(0)} = \exp(\beta_1) \end{aligned}$$

- ハザード比からは基準ハザード関数が消える

ハザード比とその解釈

- ハザード関数は死亡率の関数だったので

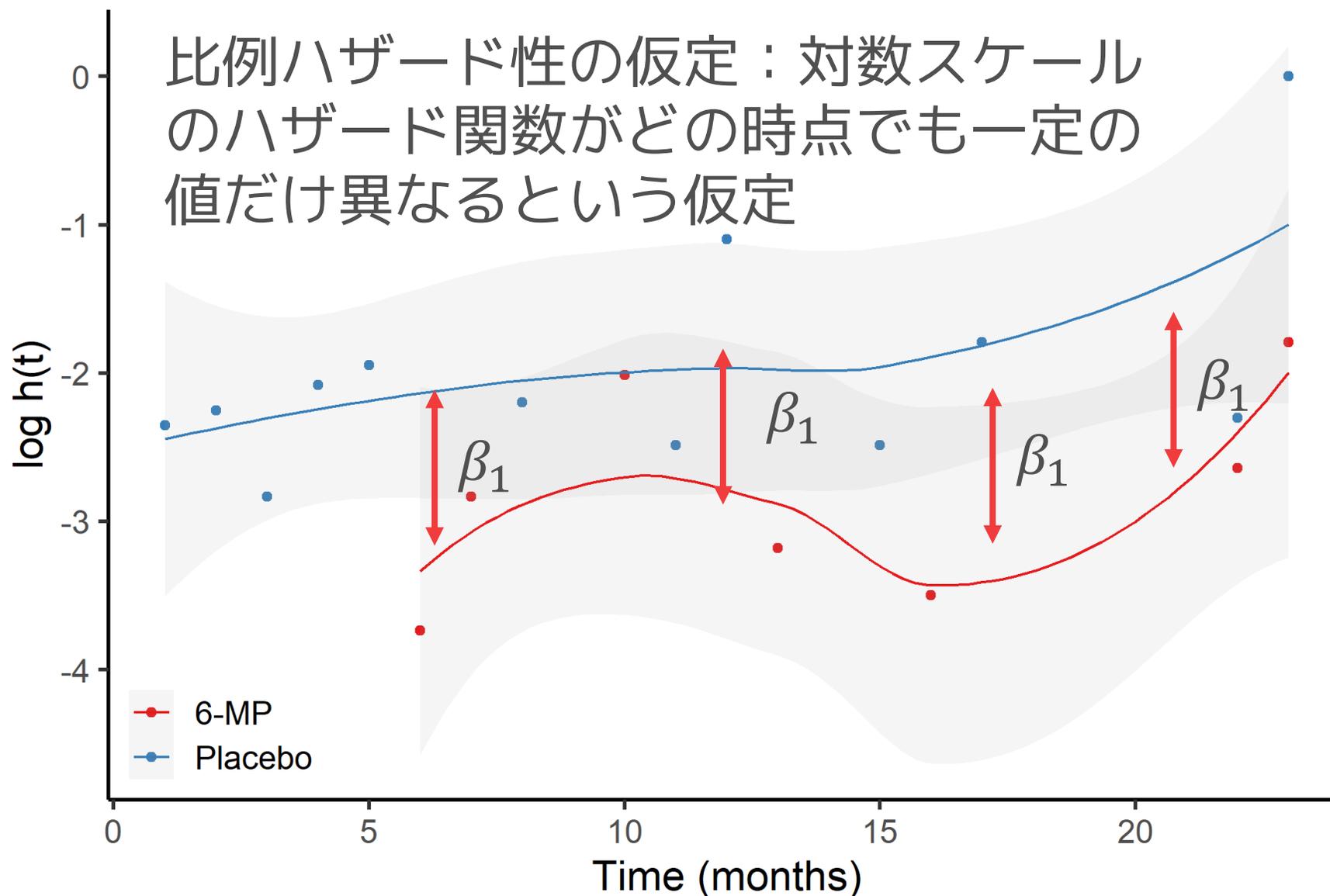
$$HR = \frac{\lambda(t | X_1 = 1)}{\lambda(t | X_1 = 0)} = \exp(\beta_1)$$

- HR は死亡率の比： $X_1 = 0$ の人の死亡率と $X_1 = 1$ の人の死亡率の比
- 時間の関数ではなくなった！（基準ハザード関数消失）
- β_1 はハザード比の対数となっている： $\beta_1 = \log HR$
- ロジスティック回帰の場合と似ているが、この率は瞬間でのイベント発生確率である点に注意

比例ハザード性

- ハザード比を求めたら基準ハザード関数がなくなり、時間の関数ではなくなる：以下の仮定と同値
 - ハザード比は時間を通じて一定 (※こうなるように意図的に作られた)
- 比例ハザード性の仮定と呼ばれる
 - 比例ハザード性が成り立たない場合、推定効率が低下し、モデルの妥当性が疑われることもある
 - 比例ハザード性の成立を確認できるソフトウェアもある

比例ハザード性



説明変数が連続の場合

- 説明変数が1つで連続の場合を考える
 - X_2 を年齢とし、 $X_2 = 30$ と $X_2 = 20$ のハザード比
 - 二つのハザード関数の比を取ると

$$\begin{aligned} HR &= \frac{\lambda(t \mid X_2 = 30)}{\lambda(t \mid X_2 = 20)} = \frac{\lambda_0(t) \exp(30\beta_2)}{\lambda_0(t) \exp(20\beta_2)} \\ &= \frac{\lambda_0(t) \exp(10\beta_2)}{\lambda_0(t)} = \exp(10\beta_2) \end{aligned}$$

- 年齢が10歳上がった時のハザード比

条件を固定した調整解析

- 説明変数が複数ある場合、一部の変数を固定した、条件を揃えた推定も可能
- $X = (L, A)$ とにおいて、 $L = l$ に固定してみよう
- $A = 1$ と $A = 0$ の比較のハザード比は？

$$\begin{aligned} HR &= \frac{\lambda(t \mid L = l, A = 1)}{\lambda(t \mid L = l, A = 0)} = \frac{\lambda_0(t) \exp(\beta_L \times l + \beta_A \times 1)}{\lambda_0(t) \exp(\beta_L \times l + \beta_A \times 0)} \\ &= \exp(\beta_A) \end{aligned}$$

Cox回帰モデルでできること

- 回帰係数の推定値に基づいてハザード比を推定できる
 - リスク比やオッズ比のような相対指標
 - ハザード関数の比の解釈は瞬間イベント発生率の比
- ロジスティック回帰モデルなどの他の回帰モデルと同様に交絡の調整も可能である
 - 多くの回帰モデルは、一部の説明変数を固定したときの推定値を得るために用いることができる

ハザード比の推定

- 線形回帰モデル
 - 最小二乗法
- ロジスティック回帰モデル
 - 最尤法 (最大尤度法)
- Cox回帰モデル
 - 最大部分尤度法
- 概念的には、それぞれデータとモデルから引いた直線の誤差が最も小さくなる $\hat{\beta}$ を推定している
 - 推定方法の詳細は難しいため、書籍等を参照

白血病データへの適用

- 単変量Cox回帰モデルを適用
 - アウトカム：再発までの時間と打ち切りの有無
 - 説明変数：治療群 (6-MP vs Placebo)

白血病のRCTデータへの適用結果

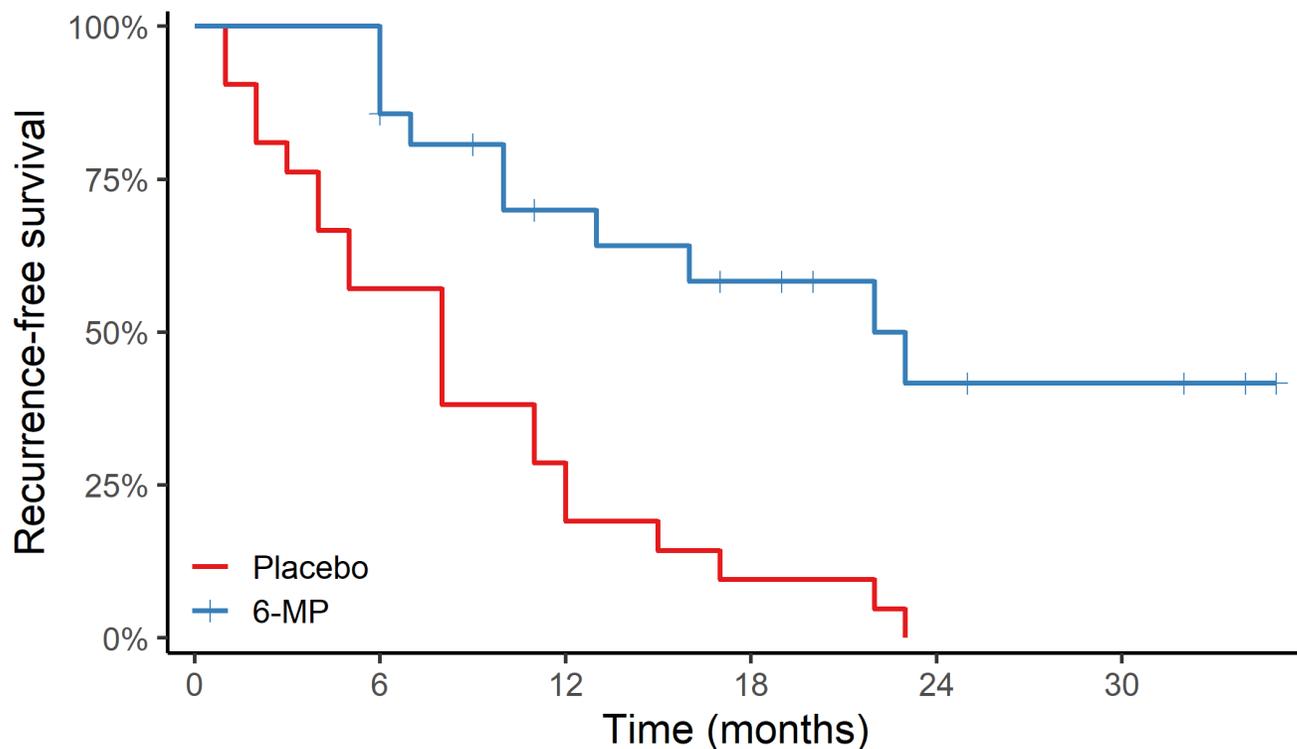
変数

HR

95%CI

P

6-MP vs. Placebo 0.234 0.107 0.509 <0.001



No. at risk		6	12	18	24	30
Placebo	21	12	6	2	0	0
6-MP	21	21	12	9	5	4

HR = 0.234 < 1であり6-MP群の方がハザードが小さい
→再発リスクが小さい

P<0.001であり、有意水準5%で統計的に有意であり、差がないという帰無仮説は棄却される

GBCSデータへの適用

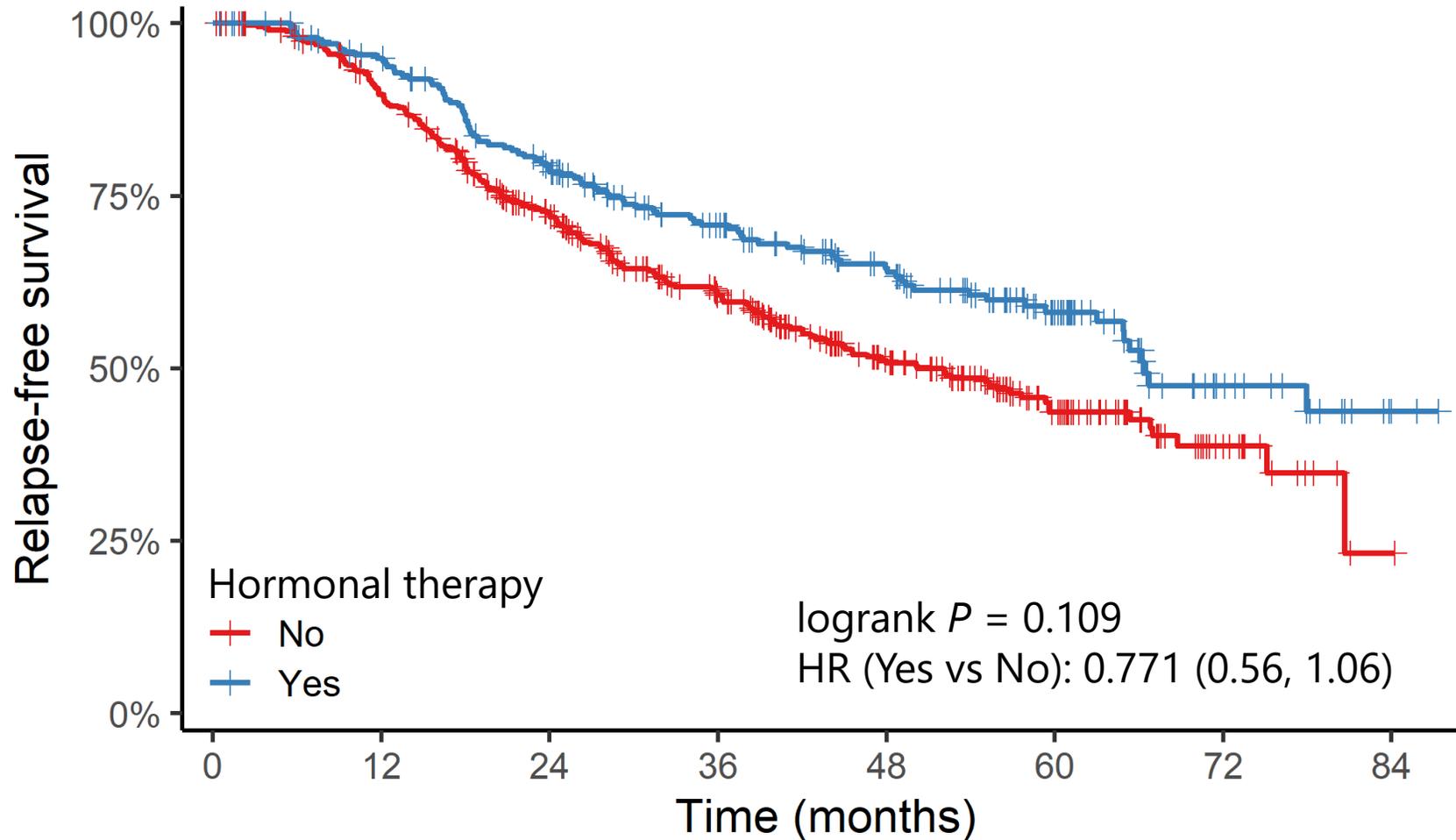
- GBCSデータ
 - 乳がんのホルモン療法の有無と再発までの時間の関係を評価した研究のデータ
- 多変量Cox回帰モデルを適用
 - アウトカム：再発までの時間と打切りの有無
 - 説明変数：ホルモン療法の有無、年齢、体重、閉経の有無

GBCSデータへの適用結果

変数	HR	95%CI	P
ホルモン療法：有 vs 無	0.681	0.529 0.876	0.002
年齢	0.986	0.969 1.004	0.125
体重	0.999	0.976 1.022	0.904
閉経：有 vs 無	1.426	0.701 2.023	0.047

- ホルモン療法と閉経の有無は統計的に有意
 - ホルモン療法有の再発リスク小、閉経有の再発リスク大
- 年齢と体重の影響は大きくない

GBCSデータへの適用結果



		No. at risk							
		0	12	24	36	48	60	72	84
No	440	379	281	195	124	63	18	1	
Yes	246	223	177	136	104	58	18	2	

解析結果の表示

- Cox回帰の場合でも、ソフトウェアのデフォルト出力をそのまま貼り付けないように
 - ハザード比などのスケールで表示 (回帰パラメータの推定値をそのまま書かない)
 - 適切な有効数字 (小数点以下2~3桁が良い)
 - Cox回帰モデルは切片はないので不要 (切片は基準ハザード関数 $\lambda_0(t)$ なので自動で推定はされない)

Cox回帰モデルの拡張

- (累積) 基準ハザード関数の推定と生存関数の推定
 - $\lambda(t) = -\frac{d}{dt} \log S(t) \Leftrightarrow \int_0^\infty \lambda(t) dt = \Lambda(t) = -\log S(t) \Leftrightarrow S(t) = \exp[-\Lambda(t)]$
 - Cox回帰の結果から $\Lambda(t)$ を推定できる (Breslow推定量)
- 説明変数が時間変化する場合にも拡張できる
 - $\lambda(t | \mathbf{X}(t)) = \lambda_0(t) \exp\{\beta_1 X_1(t) + \dots + \beta_p X_p(t)\}$
 - $\mathbf{X}(t)$: 時間変動共変量

生存時間アウトカムの回帰モデル

- 生存関数の回帰モデル
 - 加速モデル
- ハザード関数の回帰モデル
 - Cox回帰モデル (Coxの比例ハザードモデル)

加速モデルを紹介

加速モデル

- 加速時間モデルとも呼ばれる
 - 線形回帰モデルを参考にする

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- 生存時間 T は負にならない $\rightarrow T$ を対数変換

$$\log T = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

これが加速モデル

加速モデルの性質

- $\log T$ を逆変換

$$T = e^\epsilon \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

- e^ϵ は $X_1 = \cdots = X_p = 0$ のときの基準分布
- このときの生存関数を基準生存関数 $S_0(t)$ とよぶ
- T の式を生存関数のスケールにすると

$$S(t | \mathbf{X}) = S_0 \left[\frac{t}{\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)} \right]$$

加速モデルの解釈

- 説明変数が1つ (β_1 のみ) で二値の場合を考える
- 例えば $\beta_1 = 0.69$ なら

$$S(t \mid X_1 = 0) = S_0(t)$$

$$S(t \mid X_1 = 1) = S_0\left(\frac{t}{\exp(\beta_1)}\right) = S_0\left(\frac{t}{2}\right)$$

- $t = 2$ とすれば、このモデルでは $X_1 = 1$ の人の生存関数は $S_0(1)$ に等しい \rightarrow 時間が減速する

加速モデルの解釈

- 説明変数が1つ (β_1 のみ) で二値の場合を考える
- 例えば $\beta_1 = -0.69$ なら

$$S(t \mid X_1 = 0) = S_0(t)$$

$$S(t \mid X_1 = 1) = S_0\left(\frac{t}{\exp(\beta_1)}\right) = S_0(2t)$$

- $t = 2$ とすれば、このモデルでは $X_1 = 1$ の人の生存関数は $S_0(4)$ に等しい \rightarrow 時間が加速する

補足：加速モデルの導出

- $T_0 = e^\epsilon$ とおくと

$$T = T_0 \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

- これを生存関数で表すと

$$S_0(t) = \Pr(T_0 \geq t)$$

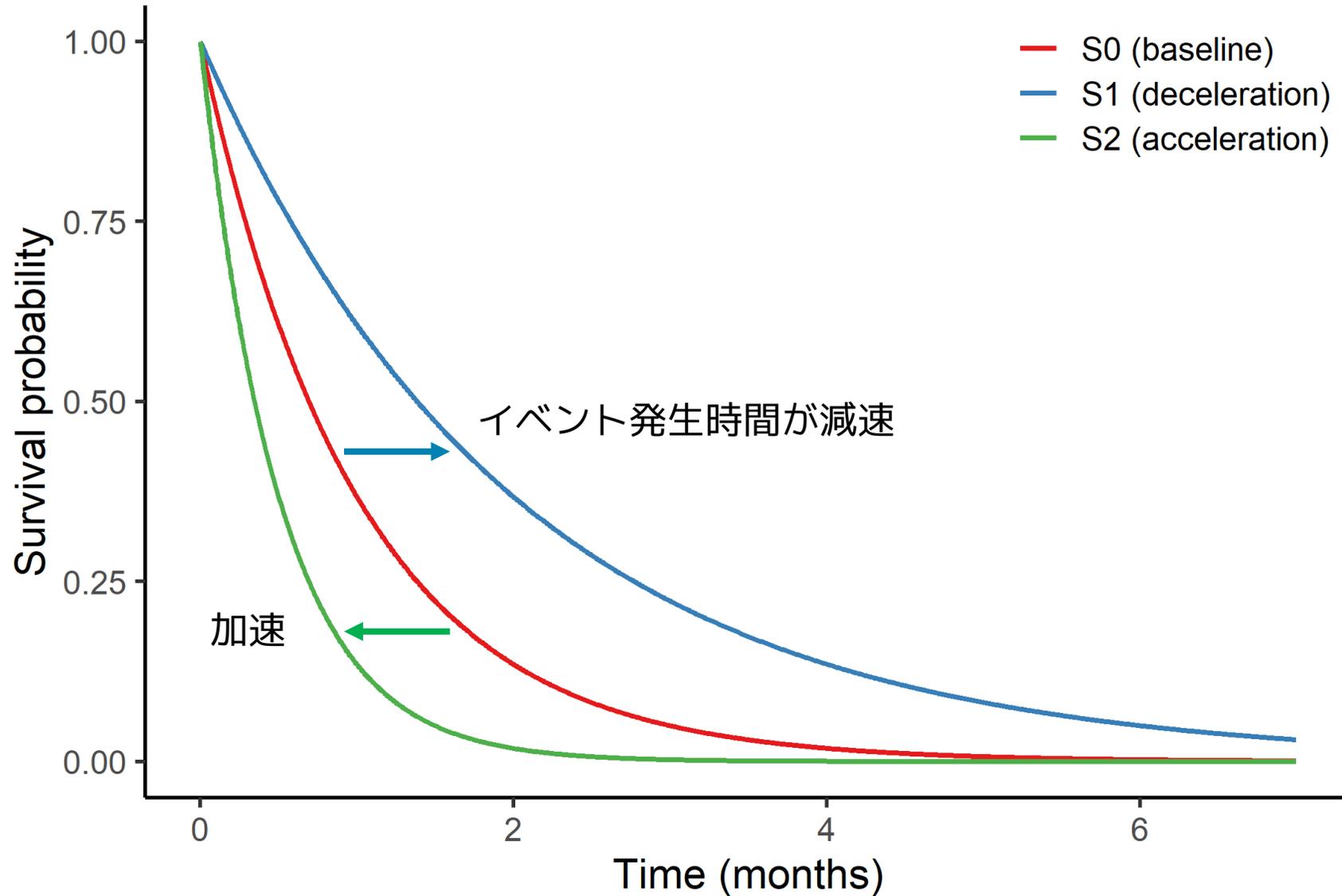
$$S(t) = \Pr(T \geq t)$$

$$= \Pr(T_0 \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \geq t)$$

$$= \Pr\left(T_0 \geq \frac{t}{\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}\right)$$

$$S(t) = S_0\left(\frac{t}{\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}\right)$$

加速モデルの図



加速モデルの結果報告

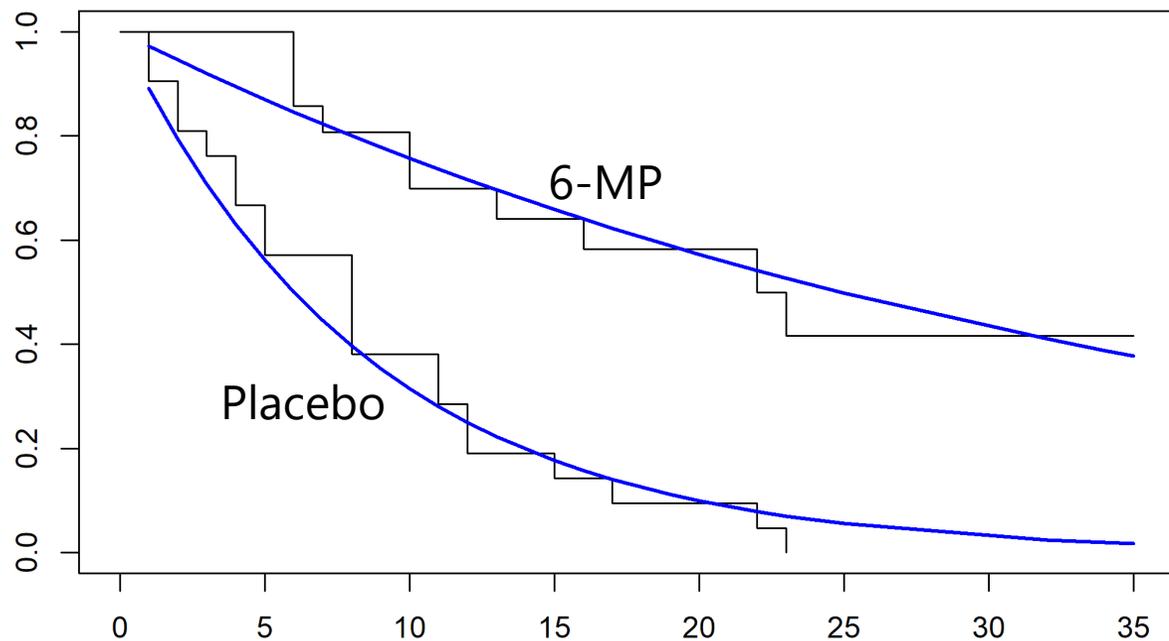
- 通常、加速因子 $\exp(\beta_1)$ を報告する
 - $S_0 \left(\frac{t}{\exp(\beta_1)} \right)$ なので
 - 加速因子 > 1 のとき：減速 (イベント減少)
 - 加速因子 < 1 のとき：加速 (イベント増加)
- ハザード比と逆になるので注意

白血病データへの適用

- 単変量の加速モデルを適用
 - アウトカム：再発までの時間と打切りの有無
 - 説明変数：治療群 (6-MP vs Placebo)

白血病のRCTデータへの適用結果

変数	加速因子	95%CI	P
6-MP vs. Placebo	4.142	1.951 8.796	<0.001



加速因子 = $4.142 > 1$ であり6-MP群のイベント時間は減速している
→再発リスクが小さい

$P < 0.001$ であり、有意水準5%で統計的に有意であり、差がないという帰無仮説は棄却される

生存時間アウトカムの回帰モデルのまとめ

- Cox回帰モデル
 - 最もよく使われる生存時間アウトカムの回帰モデル
 - 比例ハザード性が仮定できる場合強力な分析ツールとなる
- 加速モデル
 - 生存関数ベースの回帰モデル
 - 比例ハザード性の仮定は必要なし
 - それほど使われていないが、Cox回帰より少し柔軟