Rで統計解析入門

(11) 生存時間解析〔前篇〕

本日のメニュー

1. イントロ
2. イベントの無発生割合と累積発生割合の算出
3. 「イベントが起こるまでの時間」の比較
4. その他

生存時間解析とは

- ある時点から注目する事象が起きるまでの時間を解析する手法
  ⇒ 注目する事象のことを「イベント」と呼ぶ
- 生存時間解析を行う対象となる「イベント」の例
  - ガンを患っている患者さんが死亡するまでの時間
  - 臨床試験に参加している被験者が副作用を発現するまでの時間
  - システムが稼働してから故障するまでの時間
- 生存時間解析を行う対象となるデータ「イベントが起こるまでの時間」には，「イベントの有無」と「観察時間」の 2 つの変数が含まれる
  - **イベントの有無**：1 → イベントあり，0 → イベントなし
  - **観察時間**：観察を開始してから終了するまでの時間
    - イベントありの人：イベントが起こるまでの時間
    - イベントなしの人：観察を終了するまでの時間

暦日と観察期間，イベントと打ち切り

- うつ病を患っている，ある 3 人の患者さんのデータ（ ID = 1, 2, 3 ）

患者さん(ID)	イベント	観察開始日	観察終了日
1	なし	2010/4/1	2011/8/14
2	あり	2010/1/1	2011/12/2
3	なし	2010/7/1	2012/9/8

暦日と観察期間，イベントと打ち切り

- うつ病を患っている，ある 3 人の患者さんのデータ（ ID = 1, 2, 3 ）
- 生存時間解析を行う前に，データの「暦日」を「観察期間」に変換する
  - 観察開始日と観察終了日はバラバラなので，解析を行う前に各患者さんの観察開始日と観察終了日から「観察期間」を算出し，開始時点を揃える
  - 次に「観察期間」の小さい順に患者さんを並べ替える

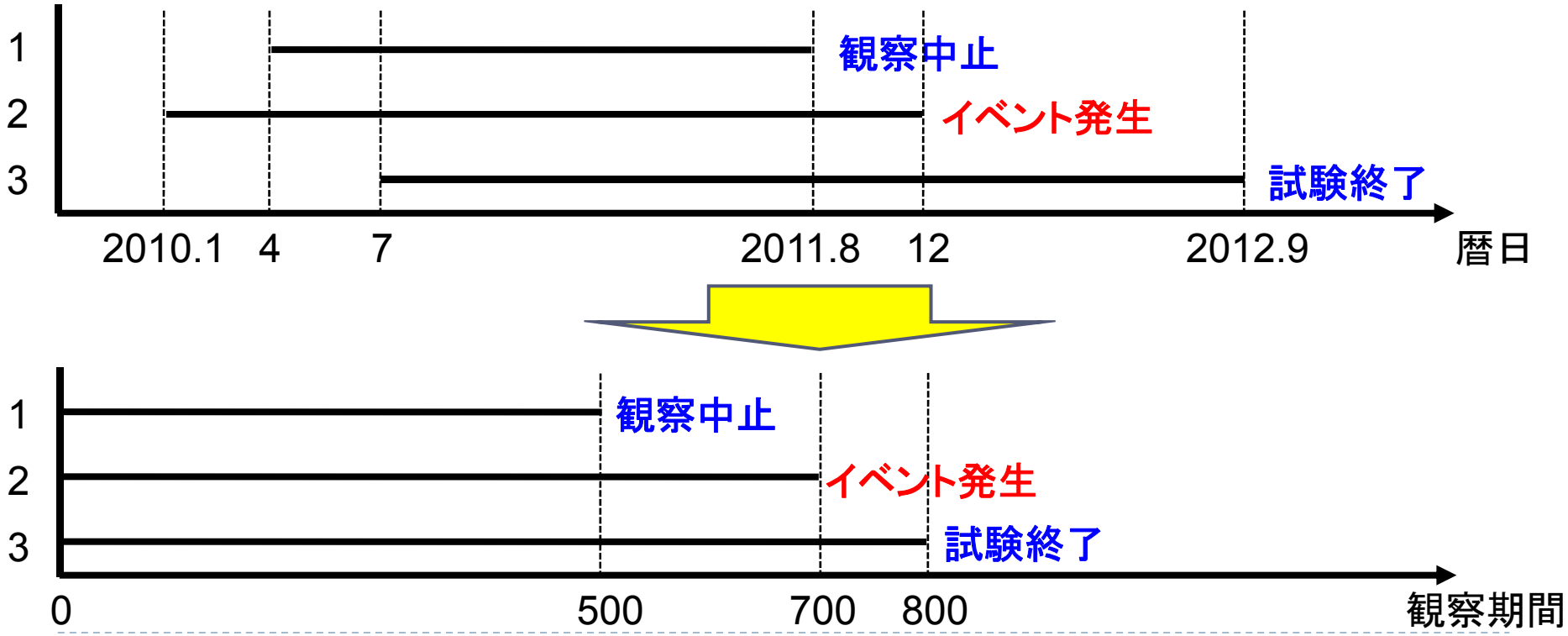
⇒ これで生存時間解析を行う準備が整う

患者さん(ID)	イベント	観察開始日	観察終了日	観察期間(日)
1	なし	2010/4/1	2011/8/14	500
2	あり	2010/1/1	2011/12/2	700
3	なし	2010/7/1	2012/9/8	800

※ 暦日：Calendar Time，観察期間：Survival Time という英語

暦日と観察期間，イベントと打ち切り

- 生存時間解析を行う前に，データの「暦日」を「観察期間」に変換する
  - 観察開始日と観察終了日はバラバラなので，解析を行う前に各患者さんの観察開始日と観察終了日から「観察期間」を算出し，開始時点を揃える
  - 次に「観察期間」の小さい順に患者さんを並べ替える



暦日と観察期間，イベントと打ち切り

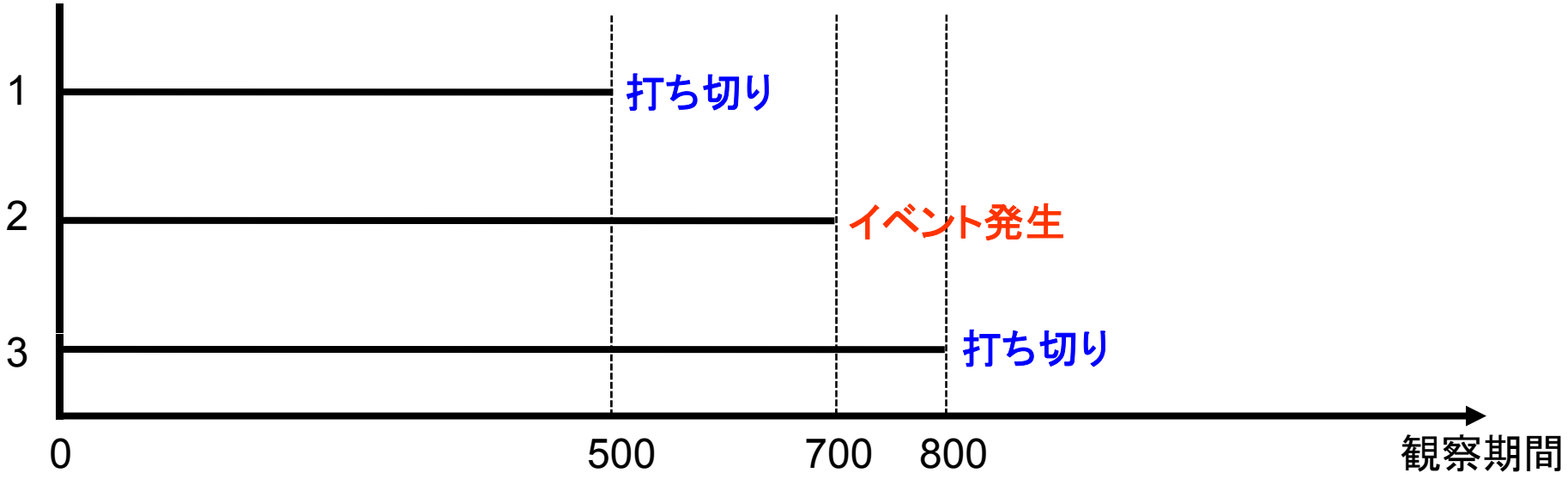
- 各患者さんの観察終了状態を見てみる
  - ID=1 の患者さん：500 日目に引っ越しにより観察中止 ⇒ 打ち切り
  - ID=2 の患者さん：700 日目に病状が改善し観察中止　⇒ イベント
  - ID=3 の患者さん：観察を続けていたが研究自体の終了時期を迎えたため，800 日目に（イベント発生も中止することもなく）観察終了 ⇒ 打ち切り
- ID=1 や ID=3 の患者さんのようにイベントが発生せずに途中で観察をやめてしまうことを「打ち切り」とよぶ
- このような患者さんのことを「打ち切り例」と呼ぶ

暦日と観察期間，イベントと打ち切り

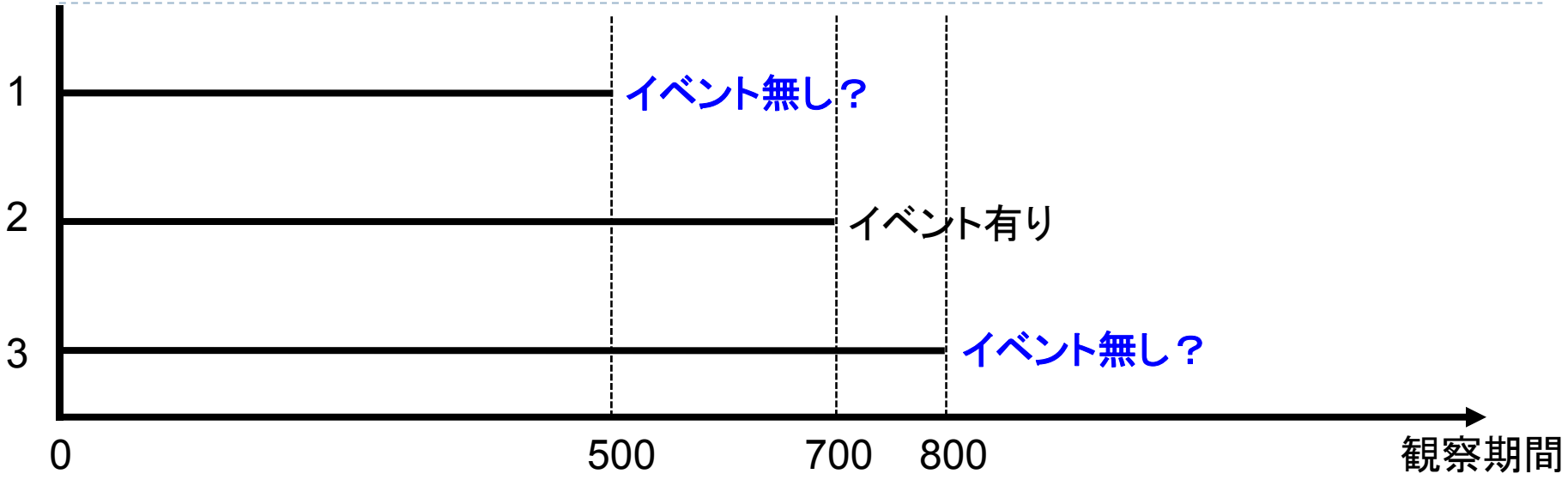
- 先ほどの観察終了状態を「打ち切り」という言葉を使って表現し直す
  - ID=1 の患者さん：500 日目に打ち切り
  - ID=2 の患者さん：700 日目にイベント
  - ID=3 の患者さん：800 日目に打ち切り
- この「打ち切り」を考慮して解析を行うことが出来ることが
  生存時間解析の特徴でありメリットである

「イベントが起きるまでの時間」について解析

- 「イベントが起きる（イベントの有無）までの時間（観察時間）」のかわりに，「イベントの有無」だけで解析，「観察時間」だけで解析すると何かまずいことがある？
- まず「観察時間」だけでを考えて解析することを考える



「観察時間」についてのみ解析

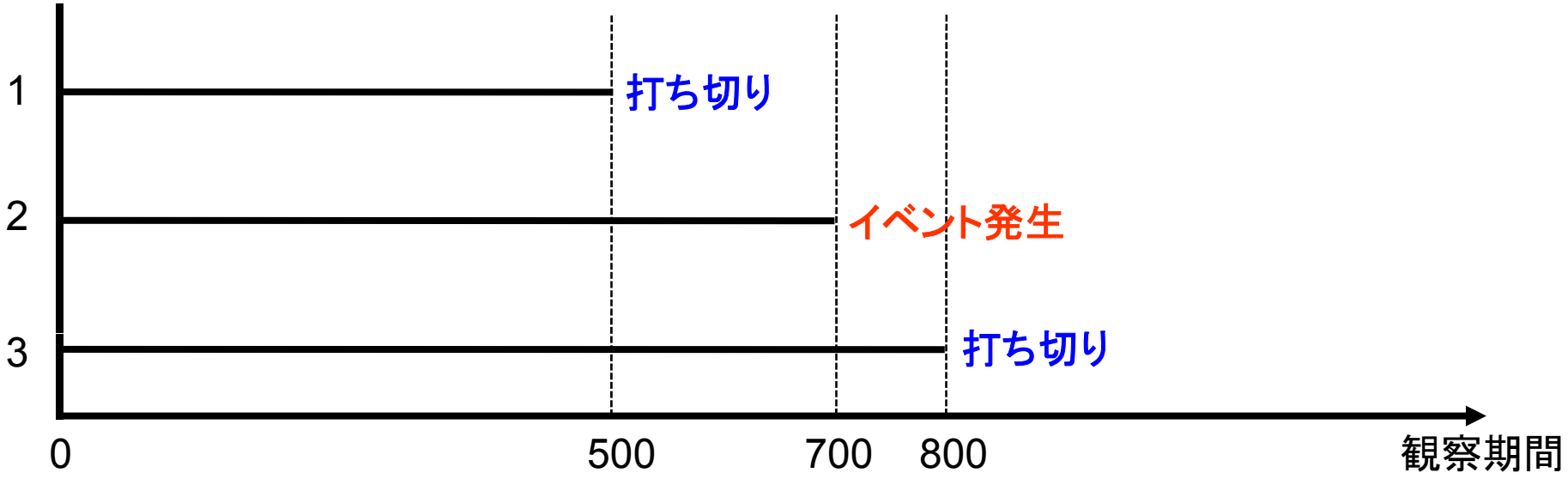


- ID=2 の 患者さん：700 日として解析に含めることになる
- ID=1 や ID=3の患者さん：イベントが発生していないため正確な観察期間は不明
  - ID=1 や ID=3 のデータを除いて解析すると「500 日目まではイベントが起きていない」や「800 日目まではイベントが起きていない」という情報が抜ける
  - ID=1やID=3 の観察期間をそれぞれと「500日」「800日」としてしまうと，「イベント有り」として解析することになるため，偏りの原因になってしまう

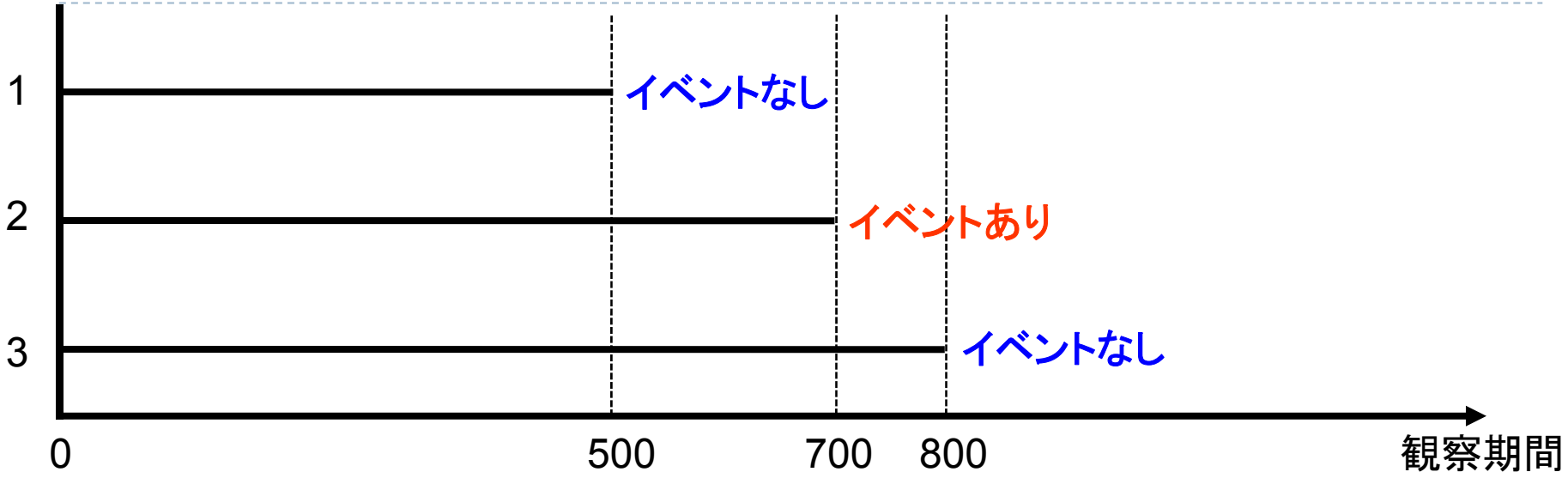
⇒ これが「打ち切り」があることの悩ましさ・・・

「イベントの有無」についてのみ解析

- 「イベントが起きる（イベントの有無）までの時間（観察時間）」のかわりに，「イベントの有無」だけで解析，「観察時間」だけで解析すると何かまずいことがある？
- 次に「イベントの有無」だけでを考えて解析することを考える

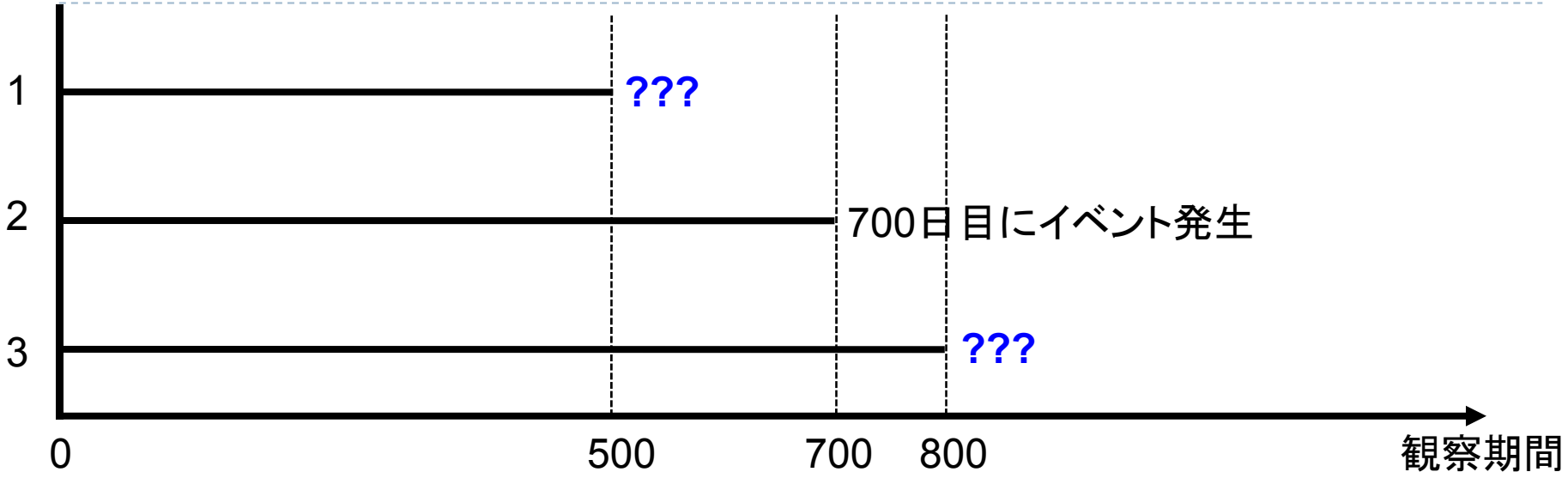


「イベントの有無」についてのみ解析



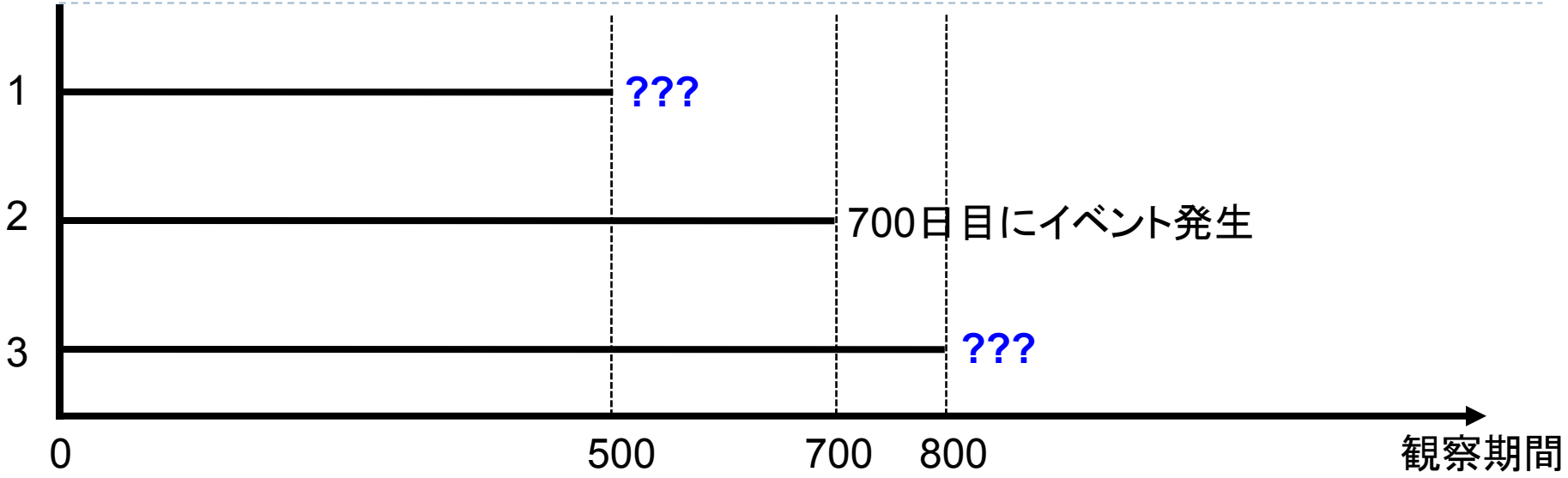
- 「イベントの有無」に関する解析なので「イベント有りの割合」を求めてみる
  - 3人中，「イベントあり」の患者さんは ID=2 の患者さん 1 人なので以下となる
    「イベントあり」の割合 ＝ 1÷3 ＝ 0.33（33.3%）
- この結果はこれで良いのだが，何となく気持ち悪さが残る
  - 患者さんの観察期間がほぼ同じならばこれでも良いという考えもある
  - 観察期間がバラバラであることはあまり気にせず，最終状態だけ気にする場合はこれでも良いという考えもある

「イベントの有無」についてのみ解析



- ID=1 の患者さんは「イベント無し」としてよい？
  - もう少し観察を続けたら ID=2 の患者さんのように「イベント有り」となるかもしれないし，「イベント無し」となるかもしれない（どちらかは不明…）
  - 仮に，「イベント有り」としてしまうと，500 日目までは「イベント無し」であった情報が抜けてしまう
  - 逆に，「イベント無し」としてしまうと，イベント発生割合を小さめにする偏りになってしまうかもしれない

「イベントの有無」についてのみ解析

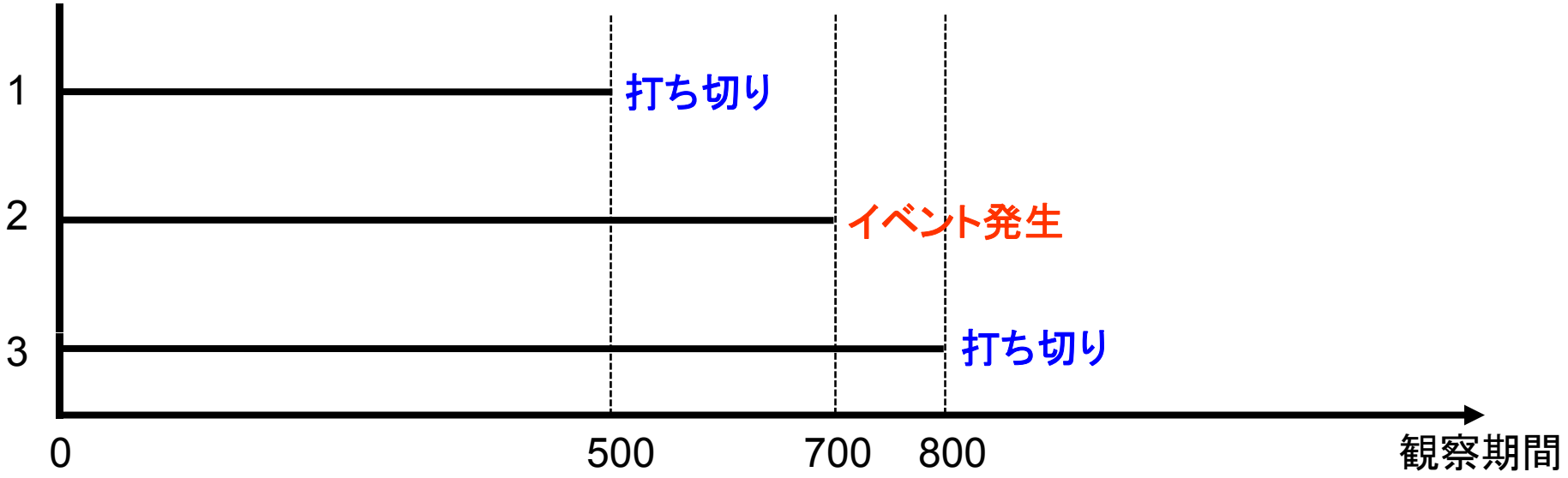


- ID=3 の患者さんは「イベント無し」としてよい？
  - 多分「イベント無し」としてよさそうですが，若干気持ち悪さが残る
  - おそらくこの気持ち悪さは各患者さんの観察期間がバラバラであることによるもの？

⇒ これ（ ID=1 や 3 のような状態）が「打ち切り」があることの悩ましさ・・・

「イベントが起きるまでの時間」について解析

- 「イベントが起きる（イベントの有無）までの時間（観察時間）」のかわりに，「イベントの有無」や「観察時間」だけで解析するとまずい場合があるので，「イベントの有無」と「観察時間」の両方を同時に解析でき，しかも「打ち切り」を考慮して解析を行うことが出来る「生存時間解析」の出番となる



本日のメニュー

1. イントロ
2. イベントの無発生割合と累積発生割合の算出
3. 「イベントが起こるまでの時間」の比較
4. その他

イベントの無発生割合の算出

- 「イベントが起きるまでの時間」を用いて，イベントが起こっていない人の割合である（イベントの無発生割合）を求めてみる
- 例として，うつ病を患っている患者さん 10 人のデータを用いる
- 「カプラン・マイヤー法」という方法により「イベントの無発生割合」を求めてみる ⇒ 「イベントの無発生割合」は「●日目の無発生割合は●である」という形で結果が出る

薬剤	イベント	時間（日）
A (ID=1)	あり	900
A (ID=2)	なし	300
A (ID=3)	あり	200
A (ID=4)	あり	800
A (ID=5)	あり	100
A (ID=6)	あり	500
A (ID=7)	あり	800
A (ID=8)	あり	700
A (ID=9)	あり	400
A (ID=10)	なし	500

イベントの無発生割合の算出

- まず最初に，時間を小さい順に並べる（日が同じ行が複数ある場合は「イベントあり」の行を上に持ってくる）

薬剤	イベント	時間（日）
A (ID=1)	あり	900
A (ID=2)	なし	300
A (ID=3)	あり	200
A (ID=4)	あり	800
A (ID=5)	あり	100
A (ID=6)	あり	500
A (ID=7)	あり	800
A (ID=8)	あり	700
A (ID=9)	あり	400
A (ID=10)	なし	500

薬剤	イベント	時間（日）
A (ID=5)	あり	100
A (ID=3)	あり	200
A (ID=2)	なし	300
A (ID=9)	あり	400
A (ID=6)	あり	500
A (ID=10)	なし	500
A (ID=8)	あり	700
A (ID=7)	あり	800
A (ID=4)	あり	800
A (ID=1)	あり	900

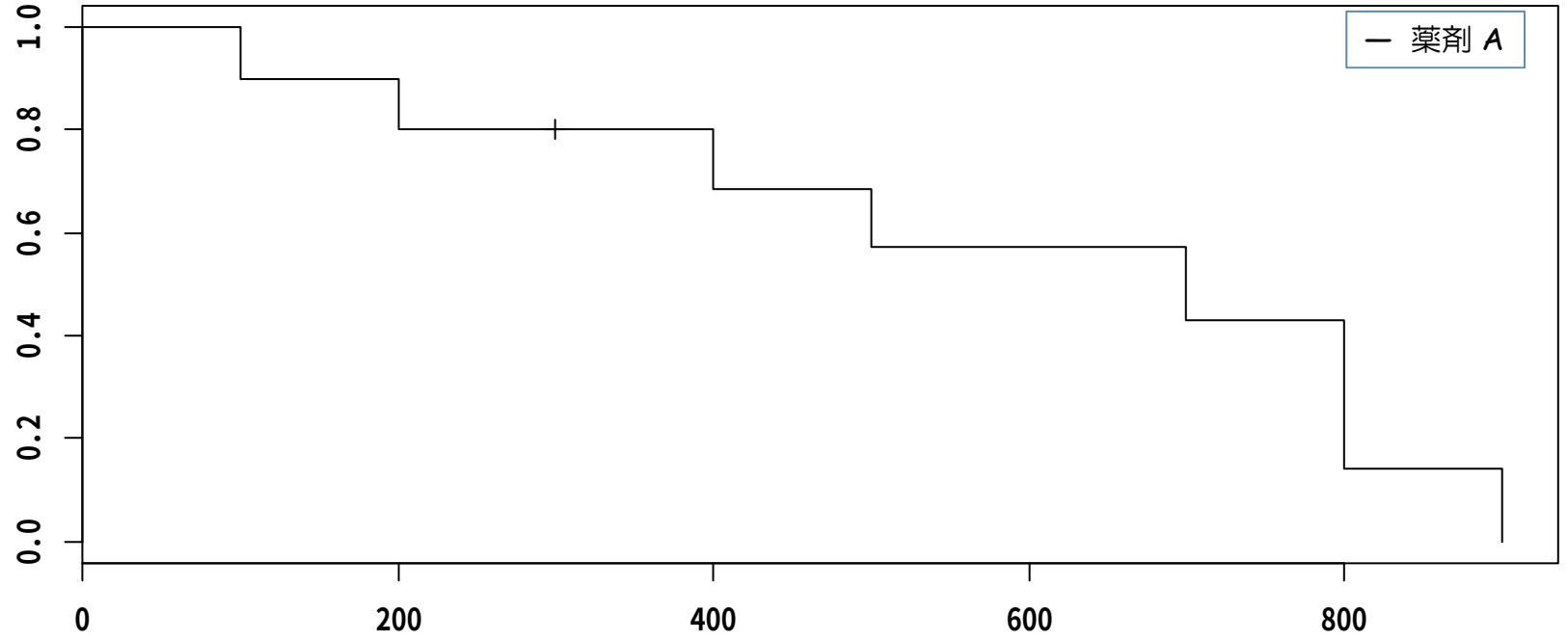
イベントの無発生割合を算出するルール

1. 0 日目の「イベントの無発生割合」を 1（100%）とする
2. 打ち切りが起こった場合は「直前までの無発生割合」をそのまま引き継ぎ，生き残っている人の数（分母，リスク集合，at risk 等の名称）を減らす
3. イベントが発生した時点で以下の計算を実行する
   「イベントの無発生割合」＝「直前までの無発生割合」
   ×「この瞬間に生き残っている人の割合」
4. 同じ日にイベントと打ち切りが起こった場合は，先にイベントが起こりその次の瞬間に打ち切りが起こったとする

- 上記ルールに従い，「イベントの無発生割合」を算出する
  - この算出方法を「カプラン・マイヤー法」という
  - 結果は「●日目の無発生割合は●である」という形となる

※ カプラン・マイヤー：Kaplan-Meier

カプラン・マイヤープロットについて



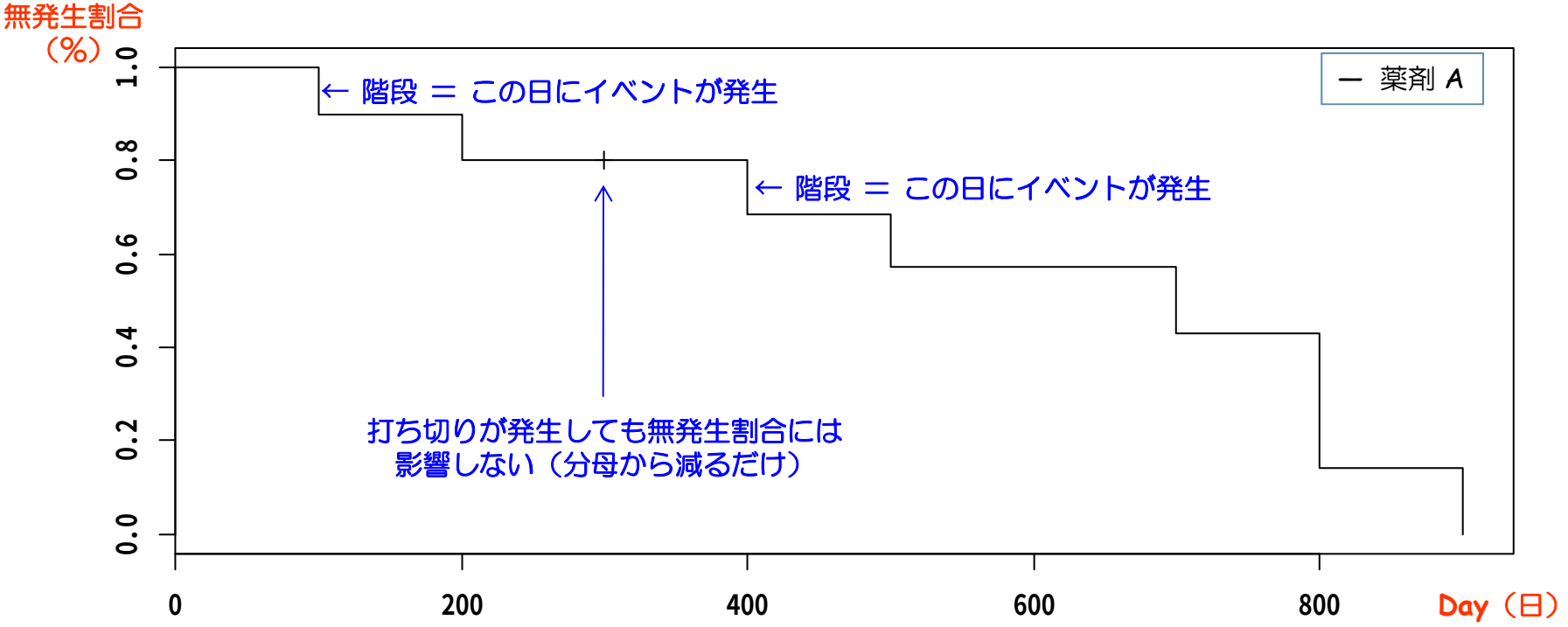
- カプラン・マイヤープロットとは，時間の経過と共にイベントの無発生割合がどのように変化するかを表したグラフ
- この曲線全体を「生存関数」のカプラン・マイヤー推定量とよんだり単に「生存曲線」ともいう

※ カプラン・マイヤー：Kaplan-Meier

前頁のグラフを描くプログラム

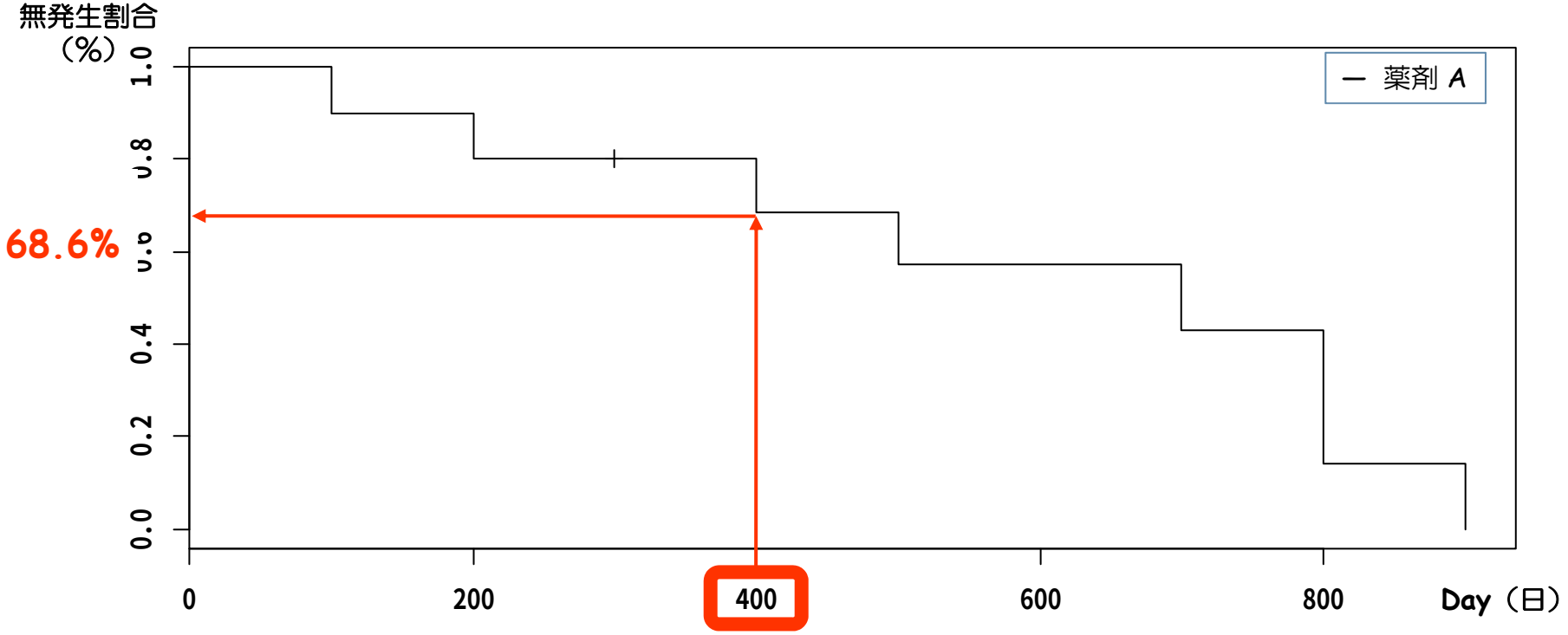
```
> library(cmprsk)
> A <- data.frame(time  =c(100,200,300,400,500,500,700,800,800,900),
+                 censor=c(  1,  1,  0,  1,  1,  0,  1,  1,  1,  1),
+                 group =rep("A",10) )
> result <- survfit(Surv(time,censor) ~ group, data=A, type="kaplan-meier")
> summary(result)

Call: survfit(formula = Surv(time, censor) ~ group, data = A,
      type = "kaplan-meier")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  100     10       1    0.900  0.0949       0.7320        1.000
  200      9       1    0.800  0.1265       0.5868        1.000
  400      7       1    0.686  0.1515       0.4447        1.000
  500      6       1    0.571  0.1638       0.3258        1.000
  700      4       1    0.429  0.1743       0.1931        0.951
  800      3       2    0.143  0.1303       0.0239        0.854
  900      1       1    0.000     NaN           NA           NA

> plot(result, conf.int=F)        # conf.int=T とすると生存関数の信頼区間を描く
```

カプラン・マイヤープロットについて



- 横軸：日
- 縦軸：無発生割合（イベントが発生していない人の割合）

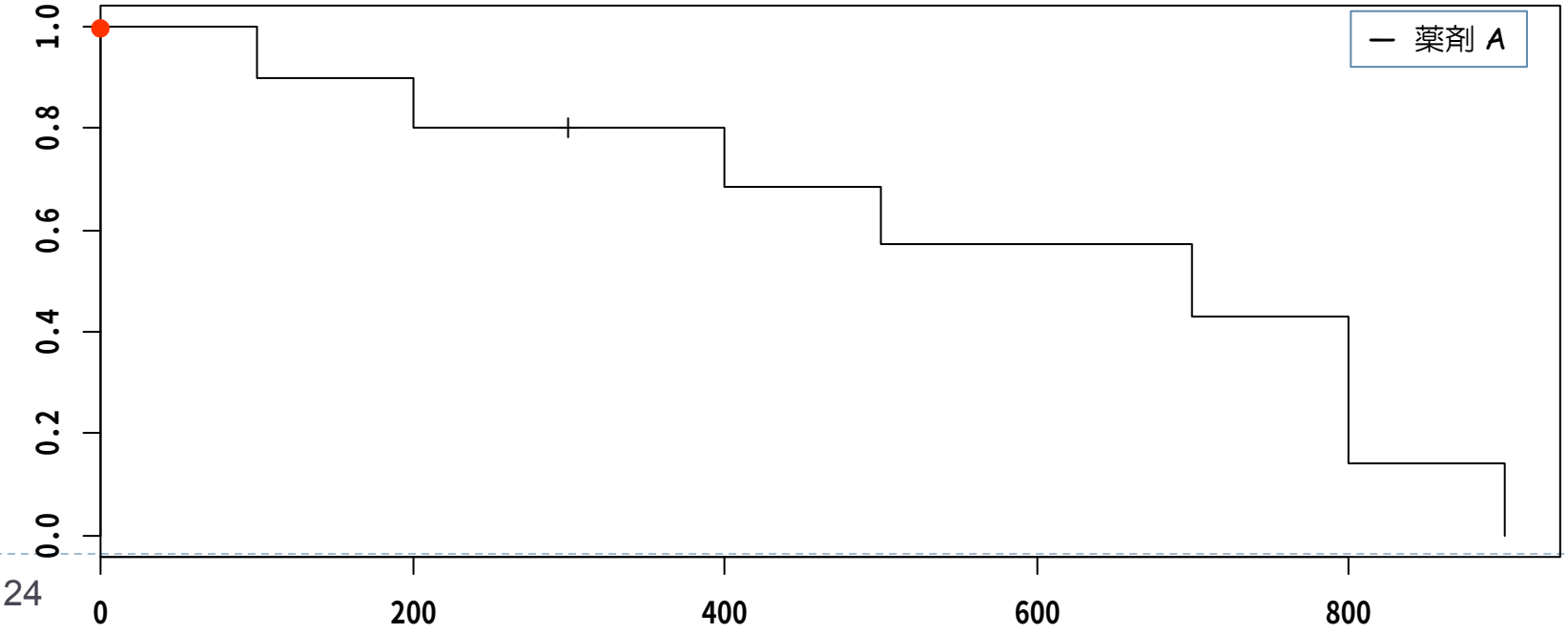
カプラン・マイヤープロットについて



- 見方の例：「400 日後の無発生割合は 68.6 %である。」
  - グラフの線が上側にある ⇒ イベント発生率が低い
  - グラフの線が下側にある ⇒ イベント発生率が高い

ルール 1 より 0 日目のイベント無発生割合は 100%

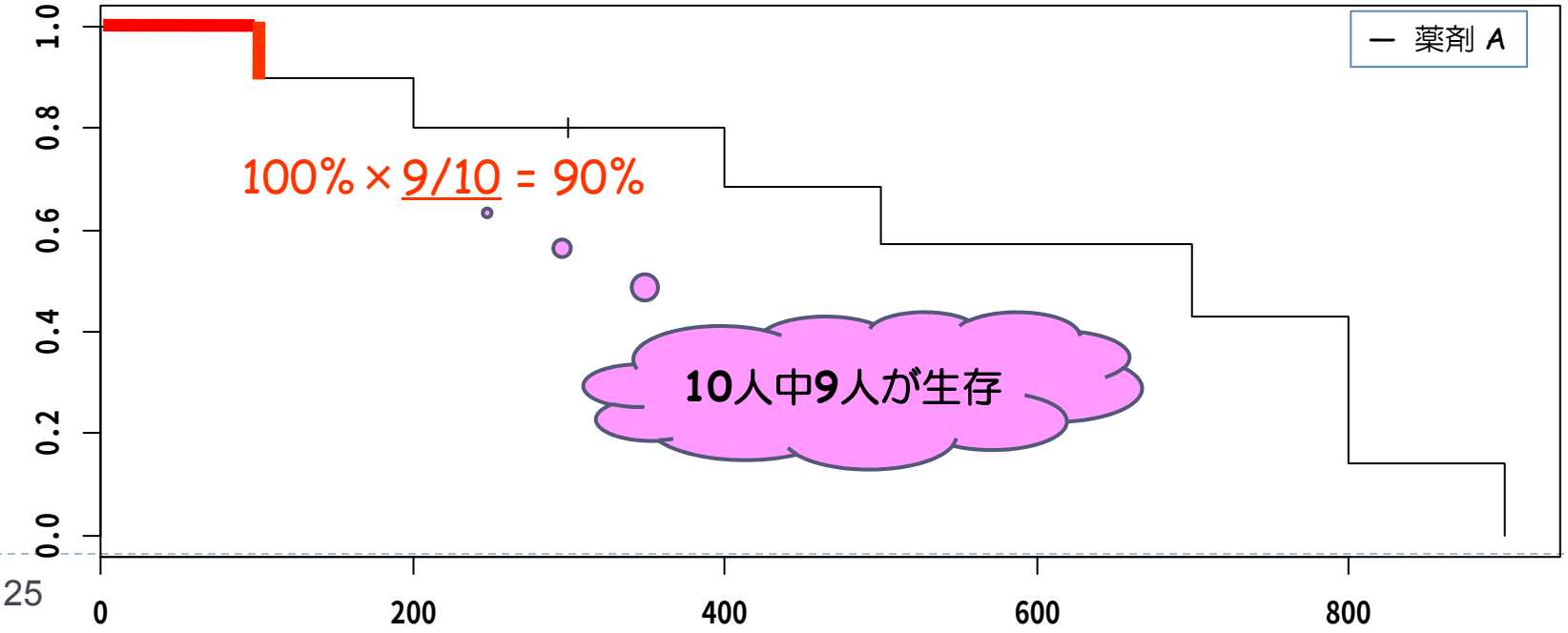
薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900



ルール 3 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

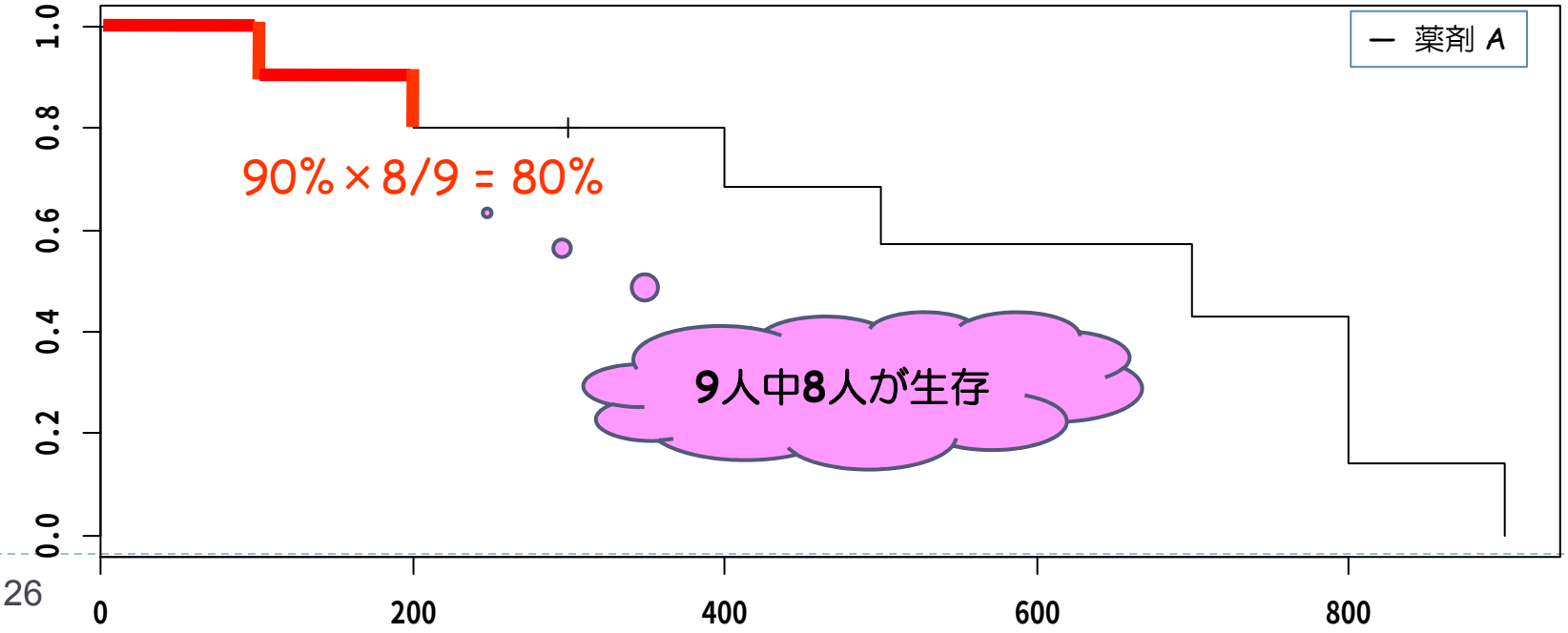
10



ルール 3 を適用

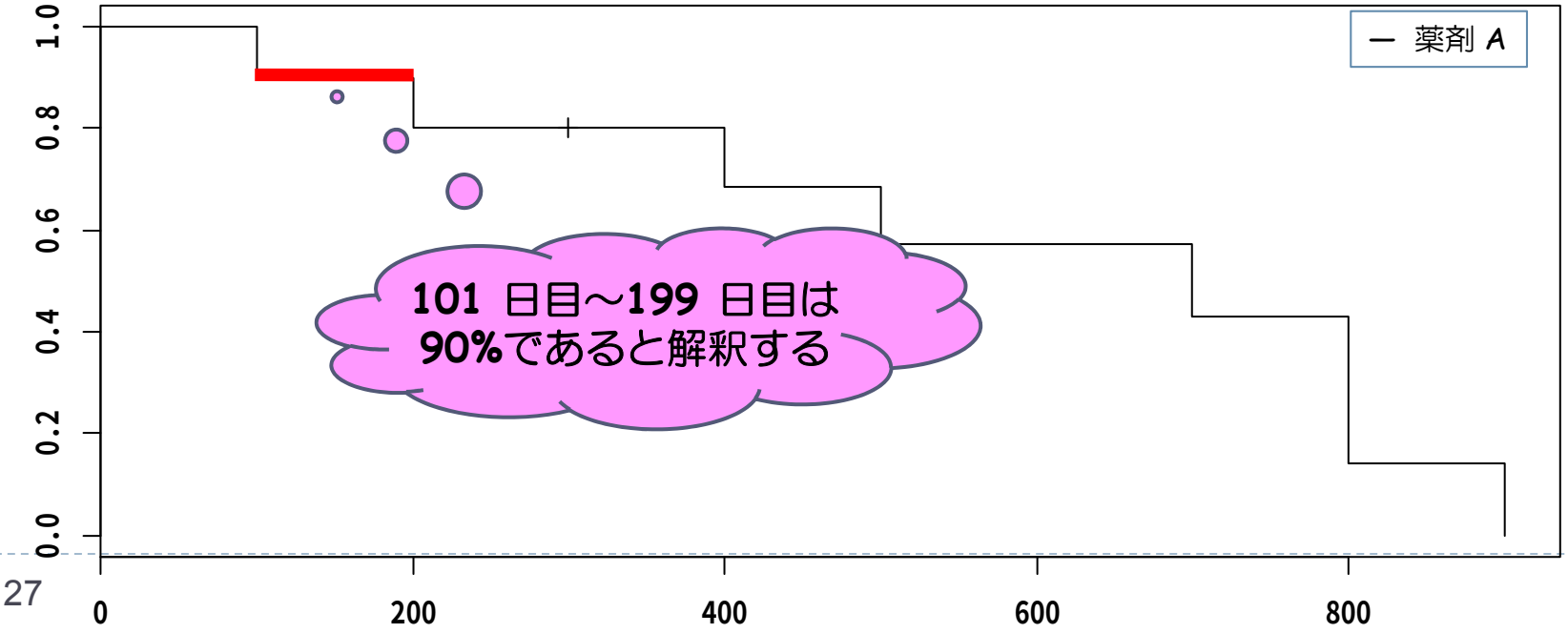
薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

9



グラフの平べったい部分の割合は・・・

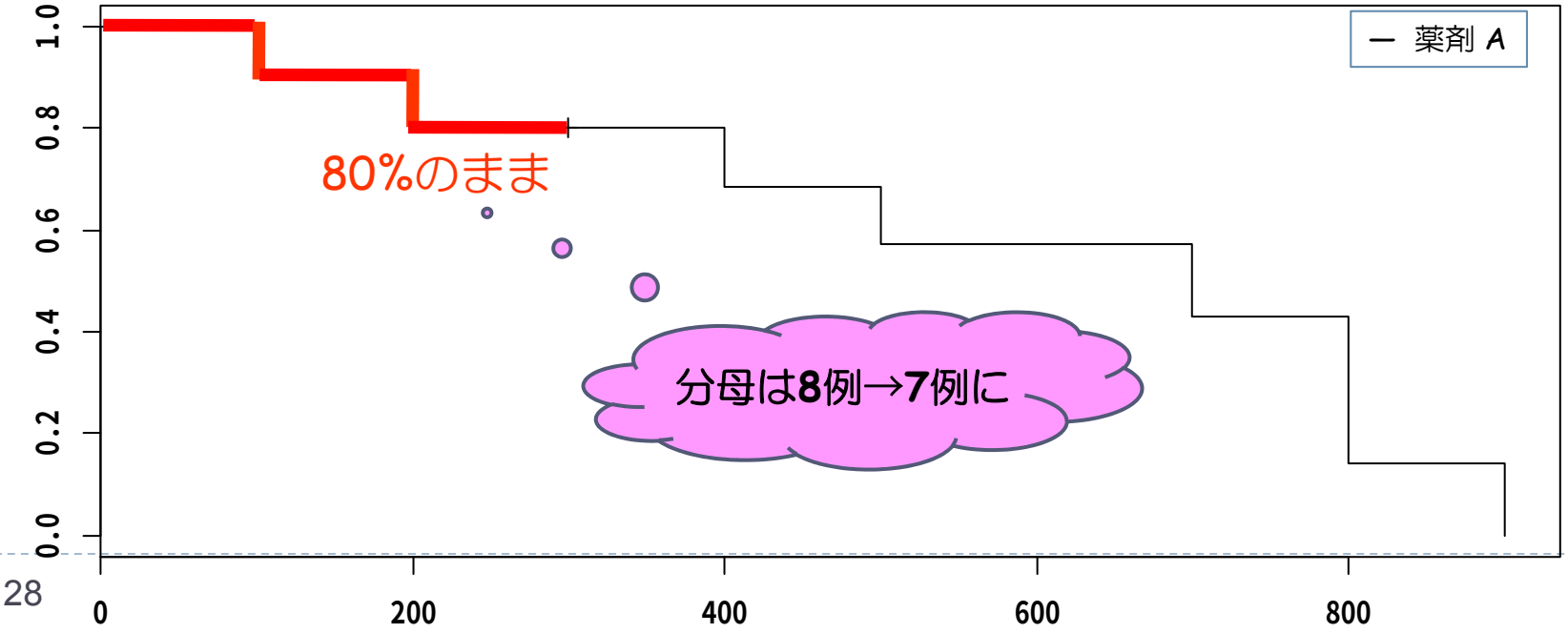
薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900



ルール 2 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

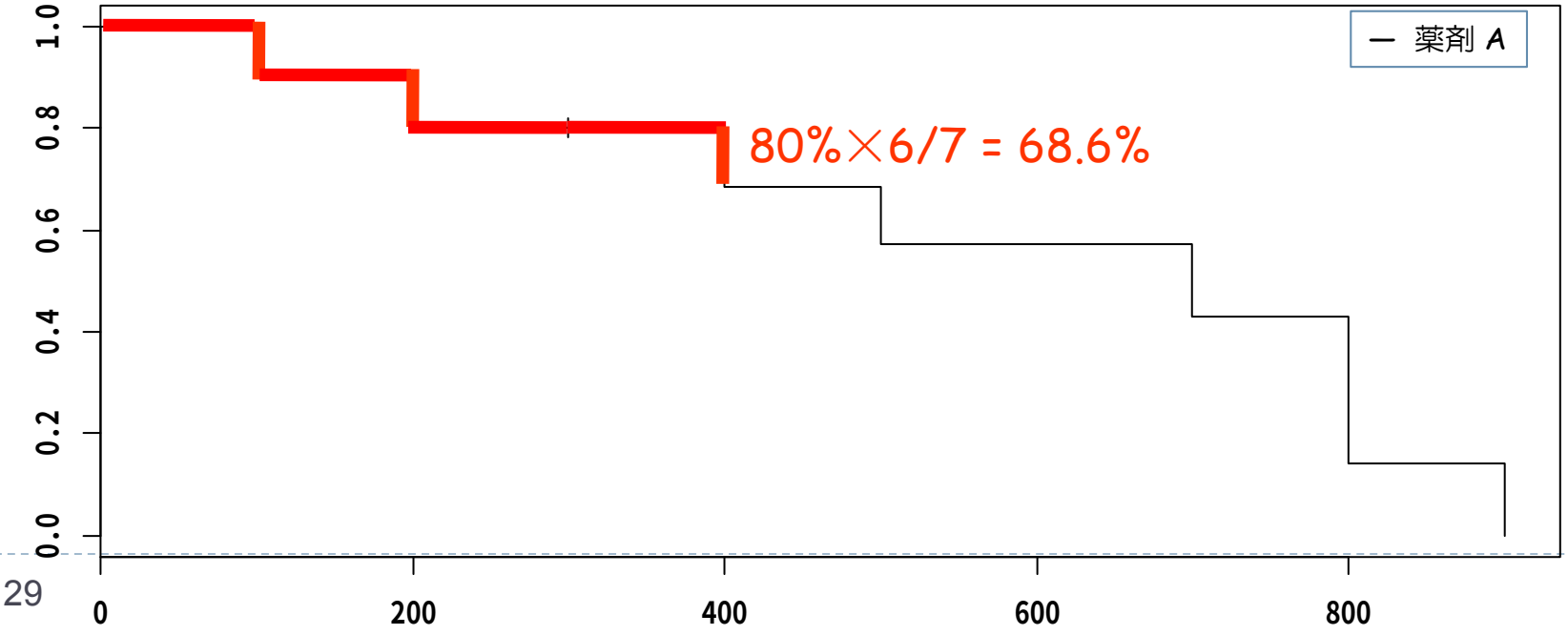
8



ルール 3 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

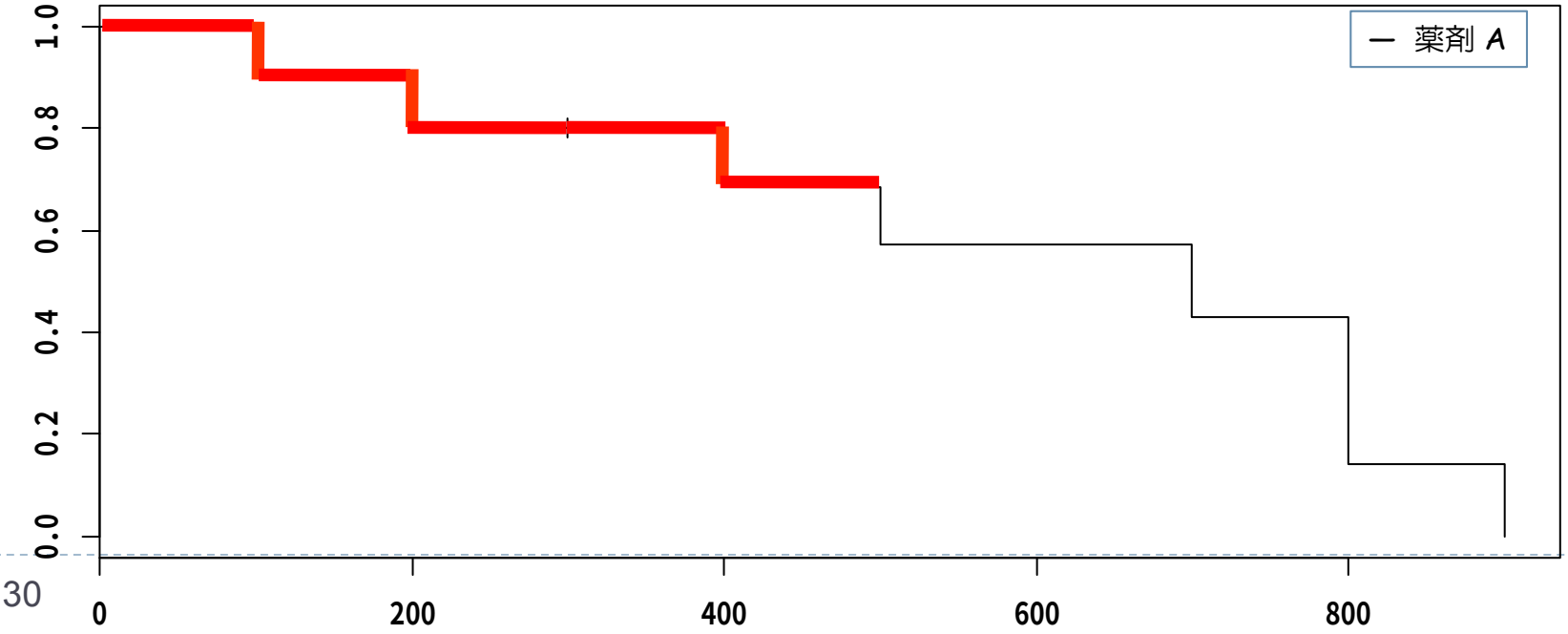
7



ルール 4 より先にイベントを考慮

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

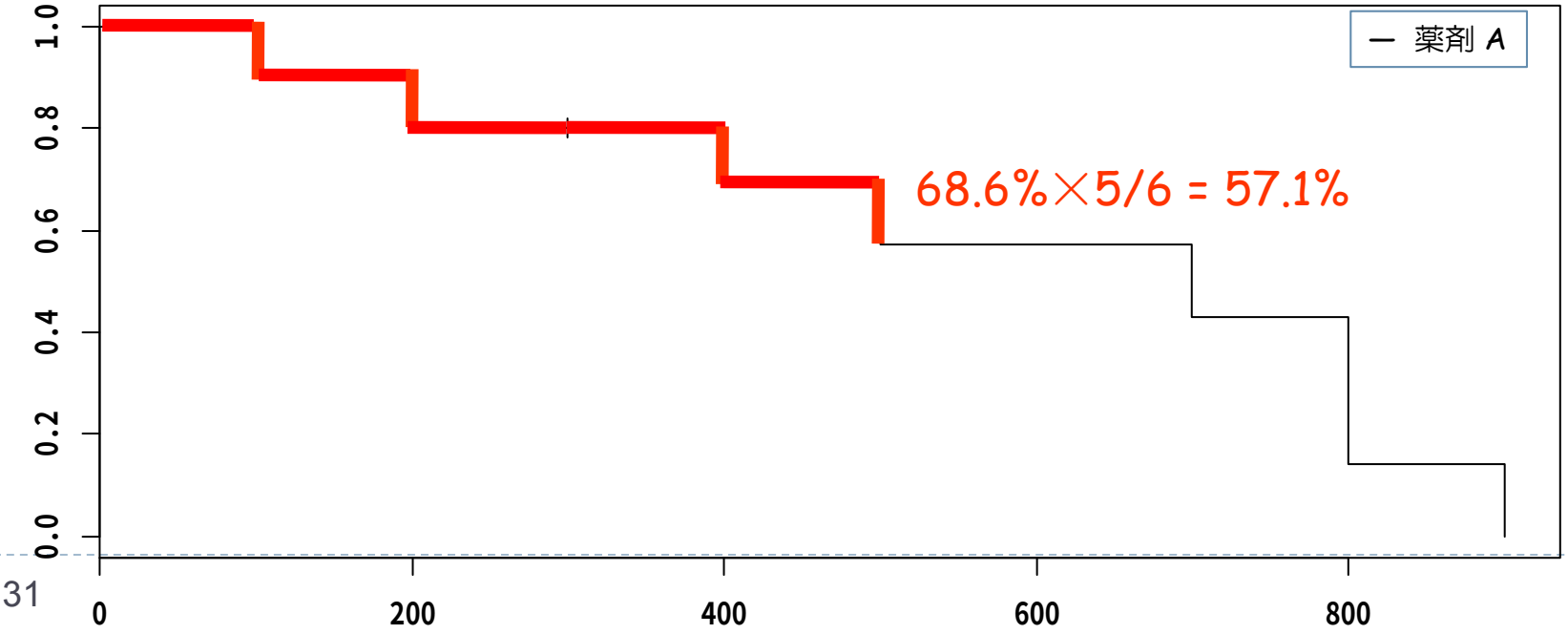
6



ルール 3 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

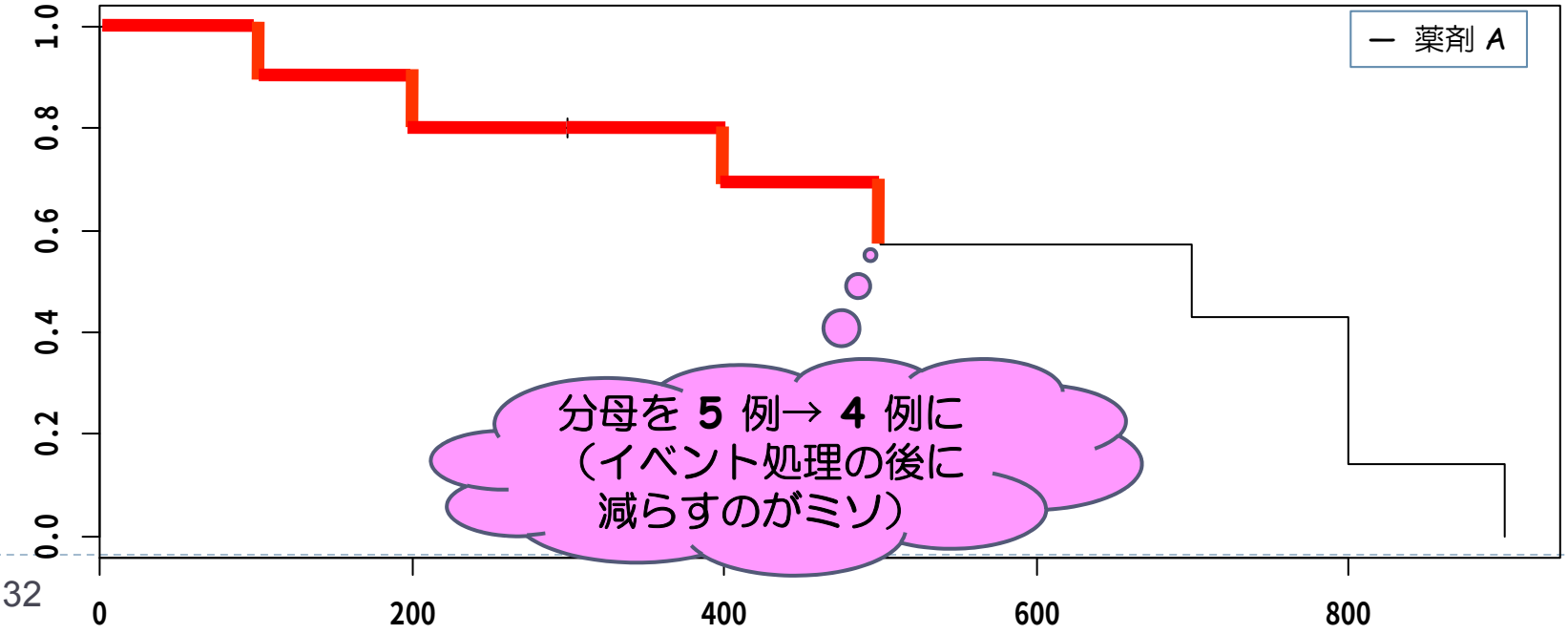
6



ルール 2 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

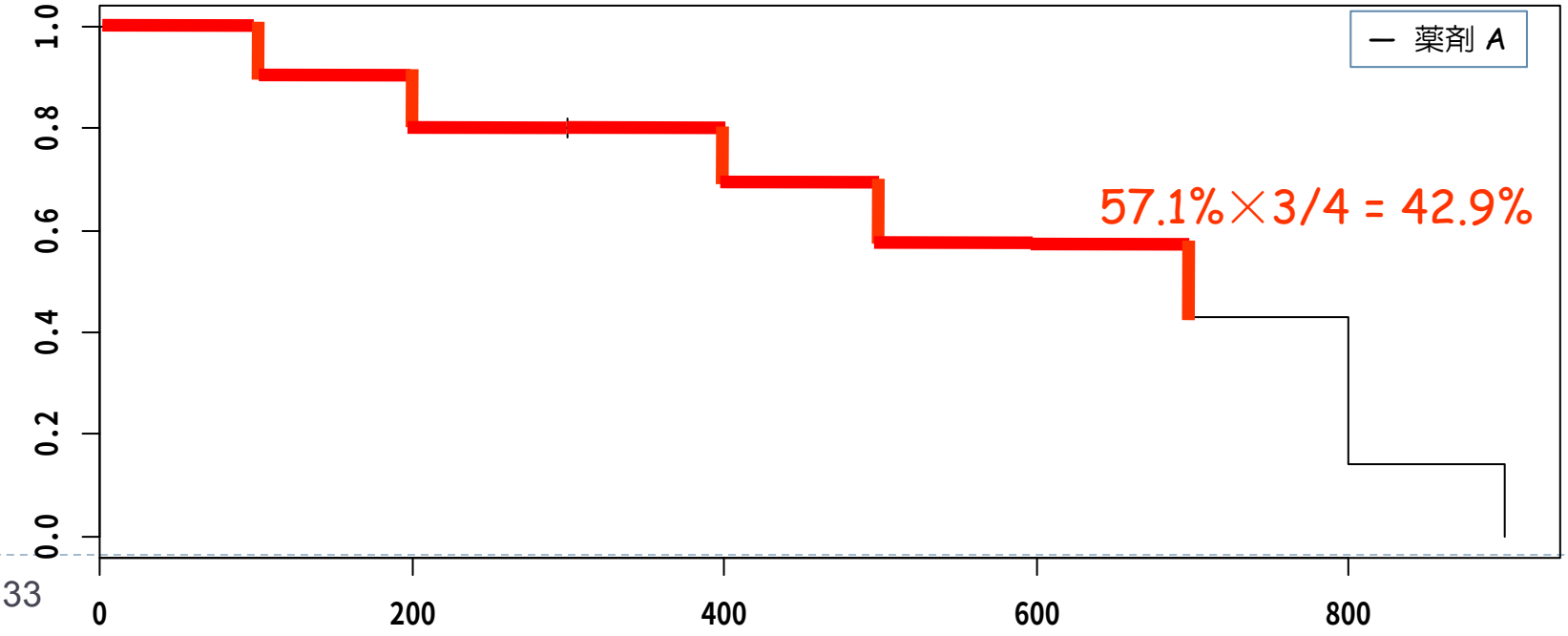
5



ルール 3 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

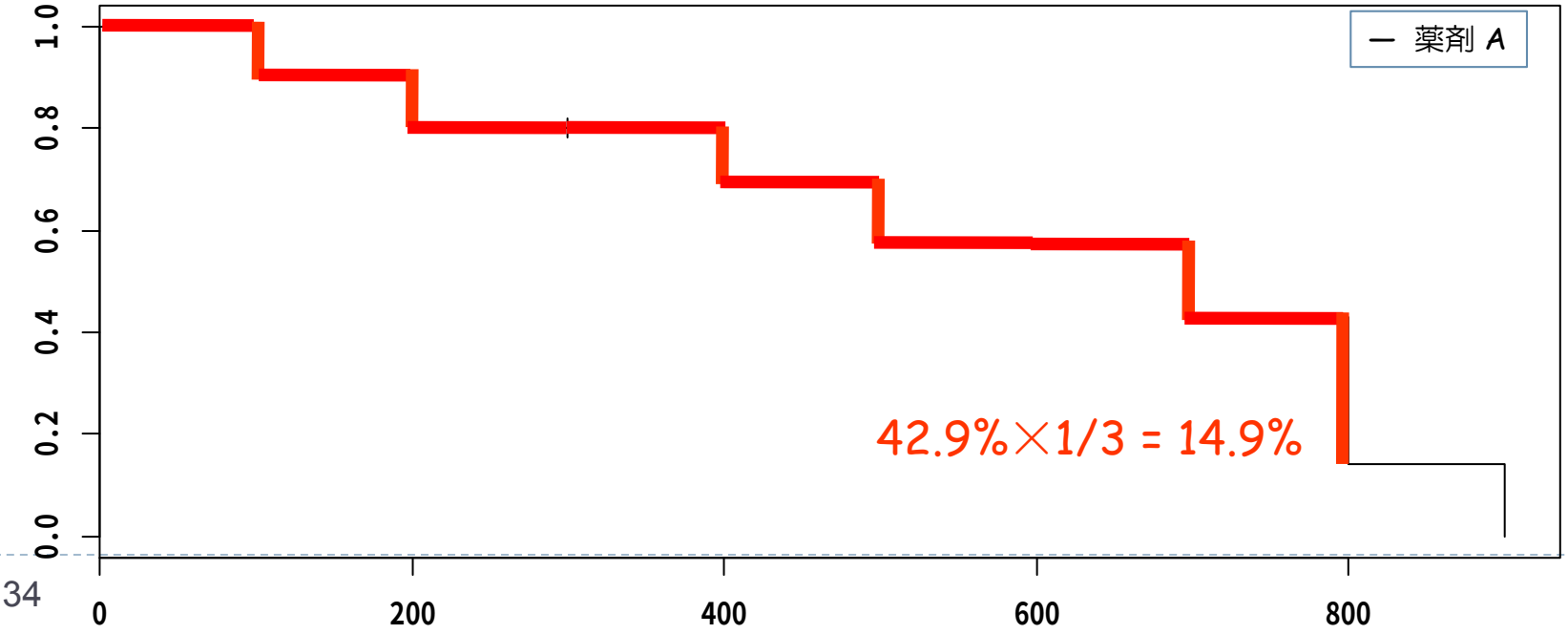
4



ルール 3 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

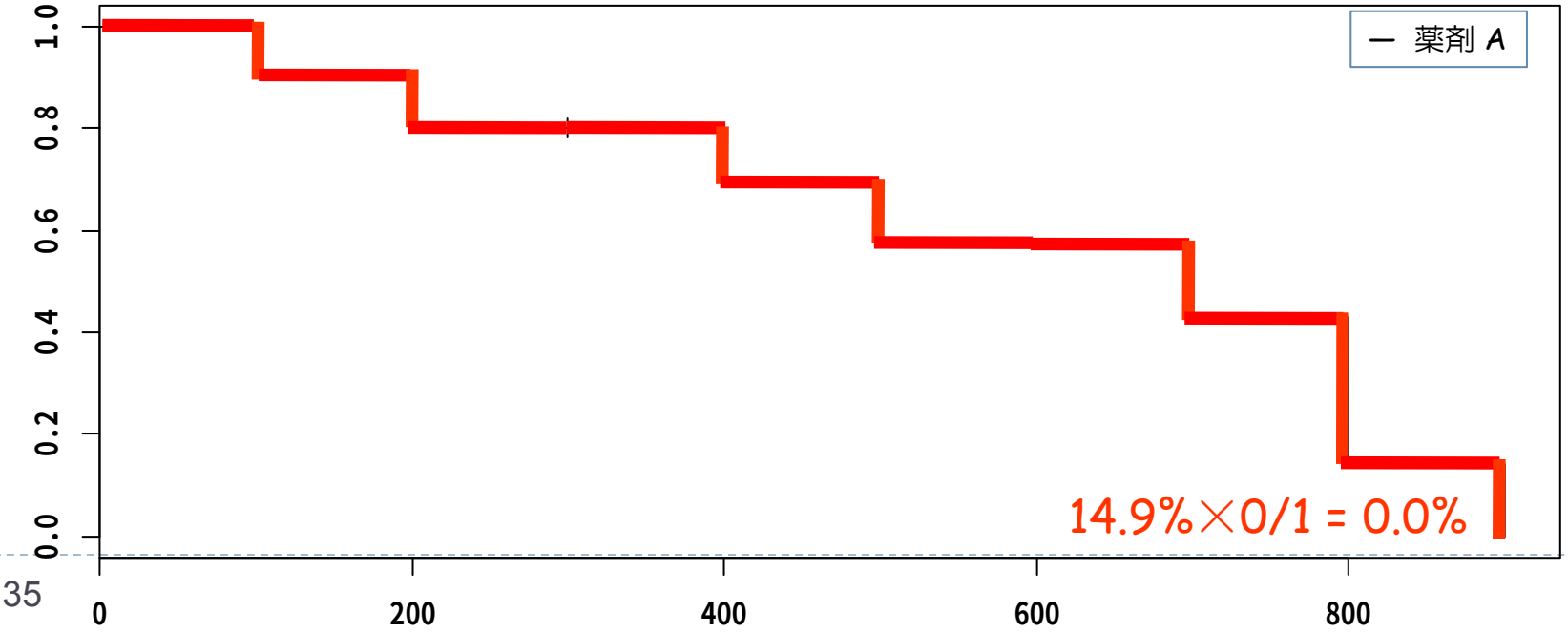
3



ルール 3 を適用

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

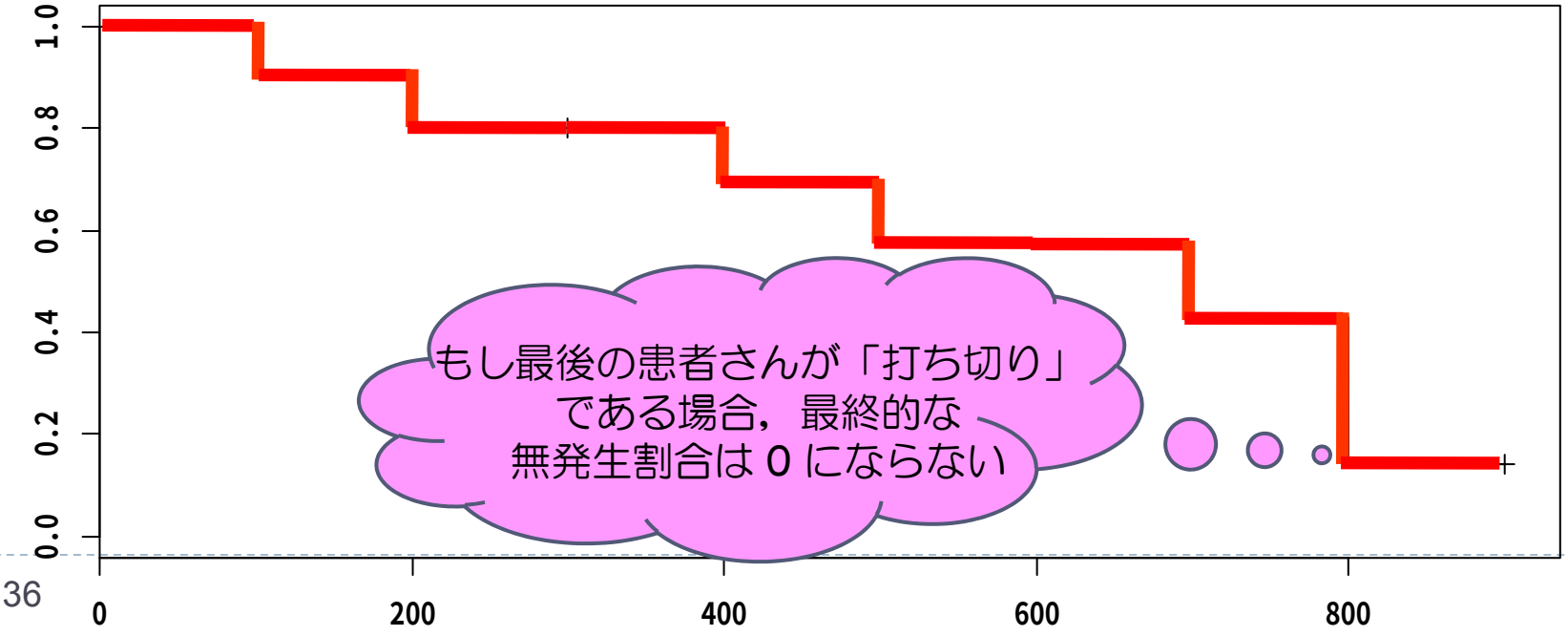
1



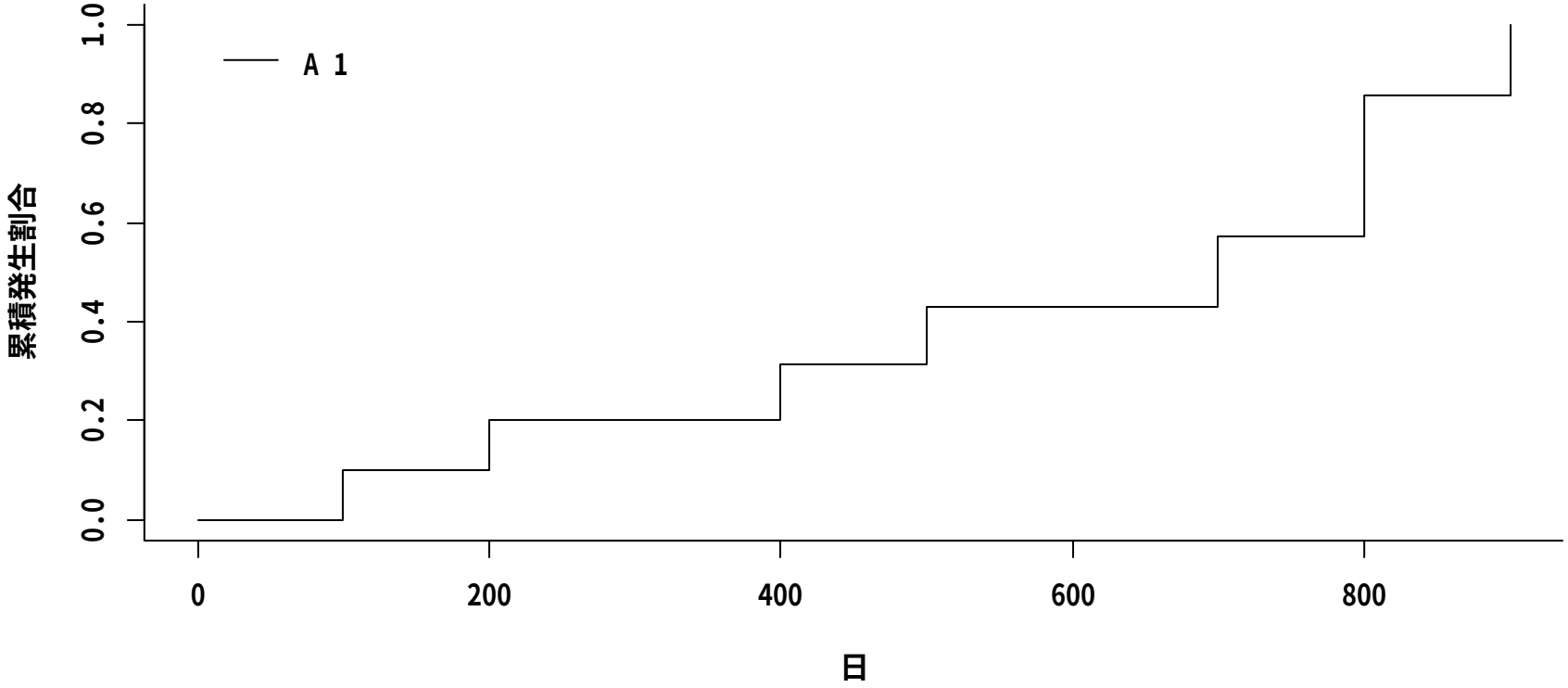
もし最後の患者さんが「打ち切り」だったら・・・

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	なし	900

1



【参考】イベントの累積発生割合



- イベントの累積発生割合を示すプロットもある
- イベントの累積発生割合：100% － イベントの無発生割合
  となるが，次々回以降の話につなげるため別の算出方法も紹介する

無発生割合と累積発生割合の関係

時間（日）	イベント	リスク集合	無発生割合	累積発生割合
0	―	10	1 (100%)	0 (0%)
100	あり	10	1×(9/10) = 0.9	0 + 1×(1/10) = 0.1
200	あり	9	0.9×(8/9) = **0.8**	0.1 + 0.9×(1/9) = 0.2
300	なし	8	**0.8**×(8/8) = 0.8	0.2 + **0.8**×(0/8) = 0.2
400	あり	7	0.8×(6/7) = 0.686	0.2 + 0.8×(1/7) = 0.314
500	あり	6	0.686×(5/6) = 0.571	0.314 + 0.686×(1/6) = 0.429
500	なし	5	0.571×(5/5) = 0.571	0.429 + 0.571×(0/5) = 0.429
700	あり	4	0.571×(3/4) = 0.429	0.429 + 0.571×(1/4) = 0.571
800	あり×2	3	0.429×(1/3) = 0.143	0.571 + 0.429×(2/3) = 0.857
900	あり	1	0.143×(0/1) = 0	0.857 + 0.143×(1/1) = 1

- 累積発生割合 = 直前の累積発生割合＋この瞬間にイベント発生した割合

  例：300 日目の累積発生割合 = 0.2 + 0.8×（イベント発生数/リスク集合）

  = 0.2 + 0.8×（0/8）= 0.2

前々頁のグラフを描くプログラム

```
> result  <- cuminc(A$time, A$censor, A$group, cencode=0)
> result2 <- timepoints(result, A$time)
> result2$"est"
    100 200 300       400       500       700       800 900
A 1 0.1 0.2 0.2 0.3142857 0.4285714 0.5714286 0.8571429   1
> plot(result, xlab="日", ylab=" (%) ")
```

準備：データ「DEP」の読み込み

1. データ「DEP」を以下からダウンロードする
   http://www.cwk.zaq.ne.jp/fkhud708/files/dep.csv
2. ダウンロードした場所を把握する ⇒ ここでは「c:/temp」とする
3. R を起動し，2. の場所に移動し，データを読み込む
4. データ「DEP」から薬剤 A と B のデータを抽出

```
> setwd("c:/temp")                          # dep.csv がある場所に移動
> getwd()                                   # 移動できたかどうか確認
> DEP <- read.csv("dep.csv")                # dep.csv を読み込む
> AB <- subset(DEP, GROUP != "C")           # 薬剤 A と B のデータを抽出
> AB$GROUP <- factor(AB$GROUP)              # 薬剤の水準を 2 カテゴリに
> AB$Y <- ifelse(AB$EVENT==1, 1, 0)         # あり→1,なし→0という変数を作成
> head(AB, n=2)
  GROUP QOL EVENT DAY PREDRUG DURATION Y
1     A  15     1  50      NO        1 1
2     A  13     1 200      NO        3 1
```

準備：架空のデータ「DEP」の変数

- GROUP：薬剤の種類（A，B，C）
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- EVENT：改善の有無（1：改善あり，2：改善なし）
  ⇒ QOL の点数が 5 点以上の場合を「改善あり（イベント発生）」とする
- Y：改善の有無（1：イベント，0：打ち切り）
  ⇒ 変数 EVENT の 2 を 0 に置き換えただけの変数
- DAY：観察期間（数値，単位は日）
- PREDRUG：前治療薬の有無（YES：他の治療薬を投与したことあり，
  NO：投与したことなし）
- DURATION：罹病期間（数値，単位は年）

準備：架空のデータ「DEP」（一部）

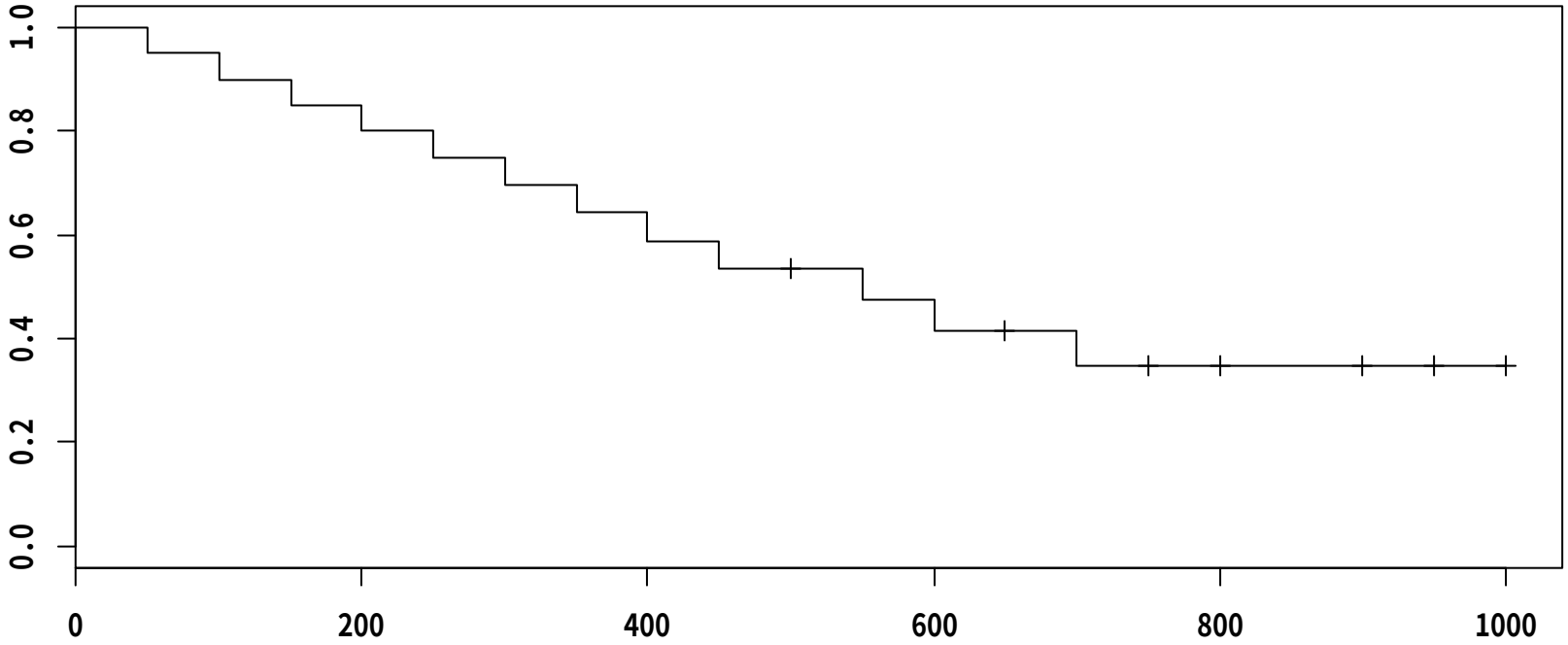
GROUP	QOL	EVENT	DAY	PREDRUG	DURATION
A	15	1	50	NO	1
A	13	1	200	NO	3
A	11	1	250	NO	2
A	11	1	300	NO	4
A	10	1	350	NO	2
A	9	1	400	NO	2
A	8	1	450	NO	4
A	8	1	550	NO	2
A	6	1	600	NO	5
A	6	1	100	NO	7
A	4	2	250	NO	4
A	3	2	500	NO	6
A	3	2	750	NO	3
A	3	2	650	NO	7
A	1	2	1000	NO	8
A	6	1	150	YES	6
A	5	1	700	YES	5
A	4	2	800	YES	7
A	2	2	900	YES	12
A	2	2	950	YES	10
B	13	1	380	NO	9
B	12	1	880	NO	5
B	11	1	940	NO	2
B	4	2	20	NO	7
B	4	2	560	NO	2
B	5	1	320	YES	11
B	5	1	940	YES	3
B	4	2	80	YES	6
B	3	2	140	YES	7
B	3	2	160	YES	13

データ「DEP」に関するイベント無発生割合

```
> result <- survfit(Surv(DAY,Y) ~ GROUP, data=AB,
+                   subset=(GROUP=="A"), type="kaplan-meier")
> summary(result)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   50     20       1    0.950  0.0487        0.859        1.000
  100     19       1    0.900  0.0671        0.778        1.000
  150     18       1    0.850  0.0798        0.707        1.000
  200     17       1    0.800  0.0894        0.643        0.996
  250     16       1    0.750  0.0968        0.582        0.966
  300     14       1    0.696  0.1037        0.520        0.932
  350     13       1    0.643  0.1087        0.462        0.895
  400     12       1    0.589  0.1120        0.406        0.855
  450     11       1    0.536  0.1139        0.353        0.813
  550      9       1    0.476  0.1158        0.296        0.767
  600      8       1    0.417  0.1156        0.242        0.718
  700      6       1    0.347  0.1153        0.181        0.666
> plot(result, conf.int=F)
```

データ「DEP」に関するイベント無発生割合

薬剤 *A* のイベント無発生割合

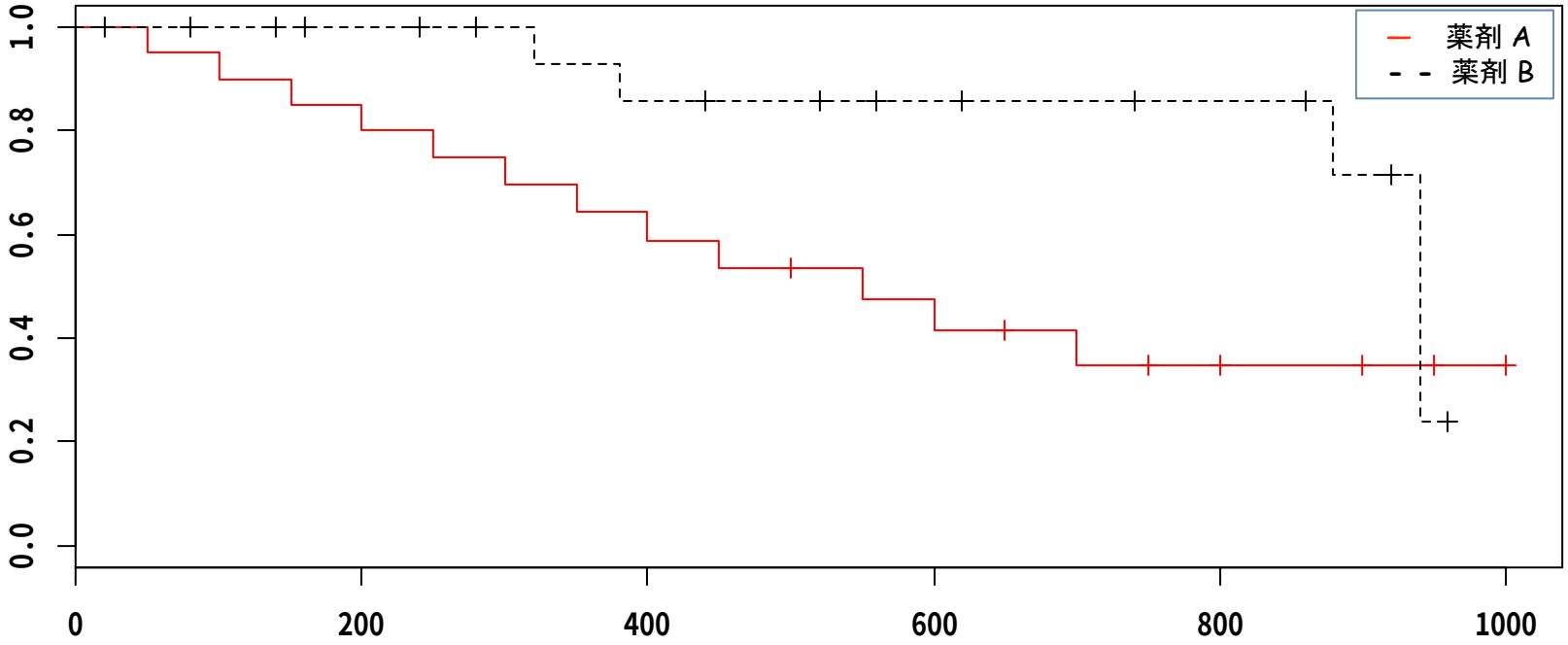


本日のメニュー

1. イントロ
2. イベントの無発生割合と累積発生割合の算出
3. 「イベントが起こるまでの時間」の比較
4. その他

各薬剤のカプラン・マイヤープロット

```
> result <- survfit(Surv(DAY,Y) ~ GROUP, data=AB, type="kaplan-meier")
> plot(result, col=2:1, lty=1:2, conf.int=F)
```



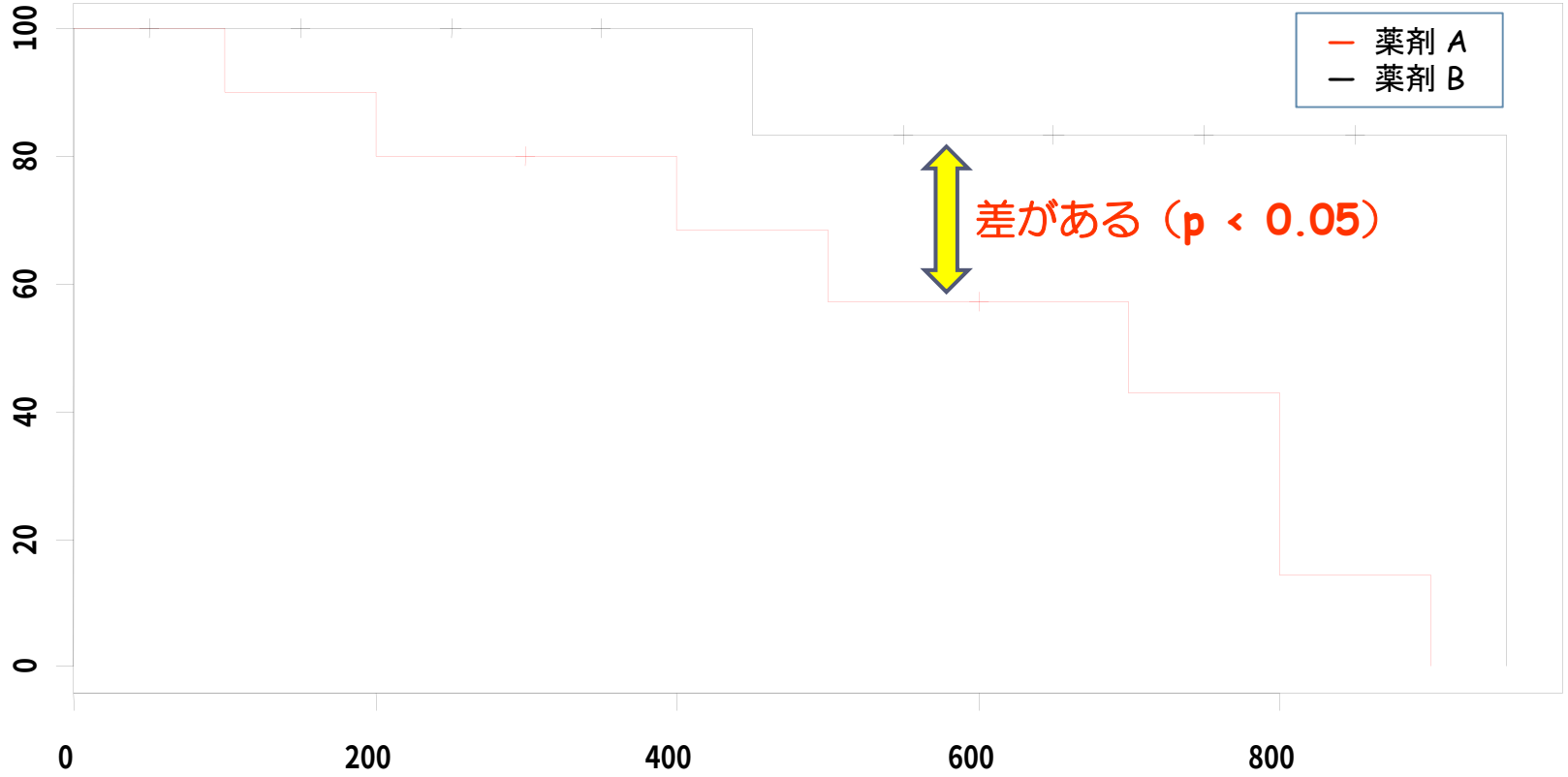
- 薬剤 A の曲線（生存関数）の方が薬剤 B よりも下
  ⇒ 薬剤 A の方がイベントの累積発生割合が高い

「イベントが起こるまでの時間」の比較

- 生存関数（生存曲線，イベントの無発生割合）の群間比較，すなわち「イベントが起こるまでの時間の群間比較」を「ログランク検定」により行う
- ログランク検定の仮説は以下：
  - 帰無仮説 H_0：各薬剤の生存関数に違いがない
  - 対立仮説 H_1：各薬剤の生存関数に違いがある

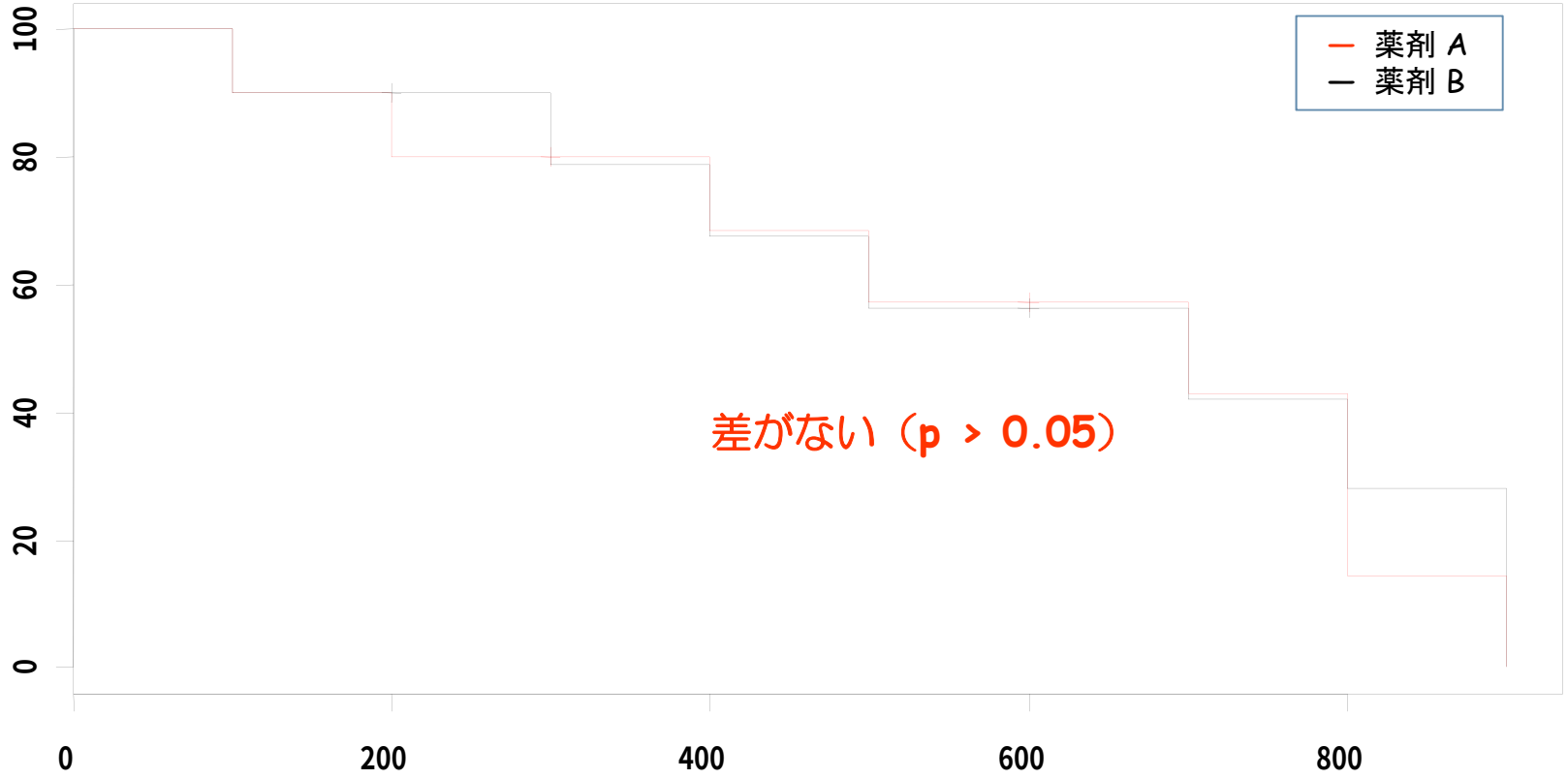
※ ログランク検定：logrank test

カプラン・マイヤープロットとログランク検定



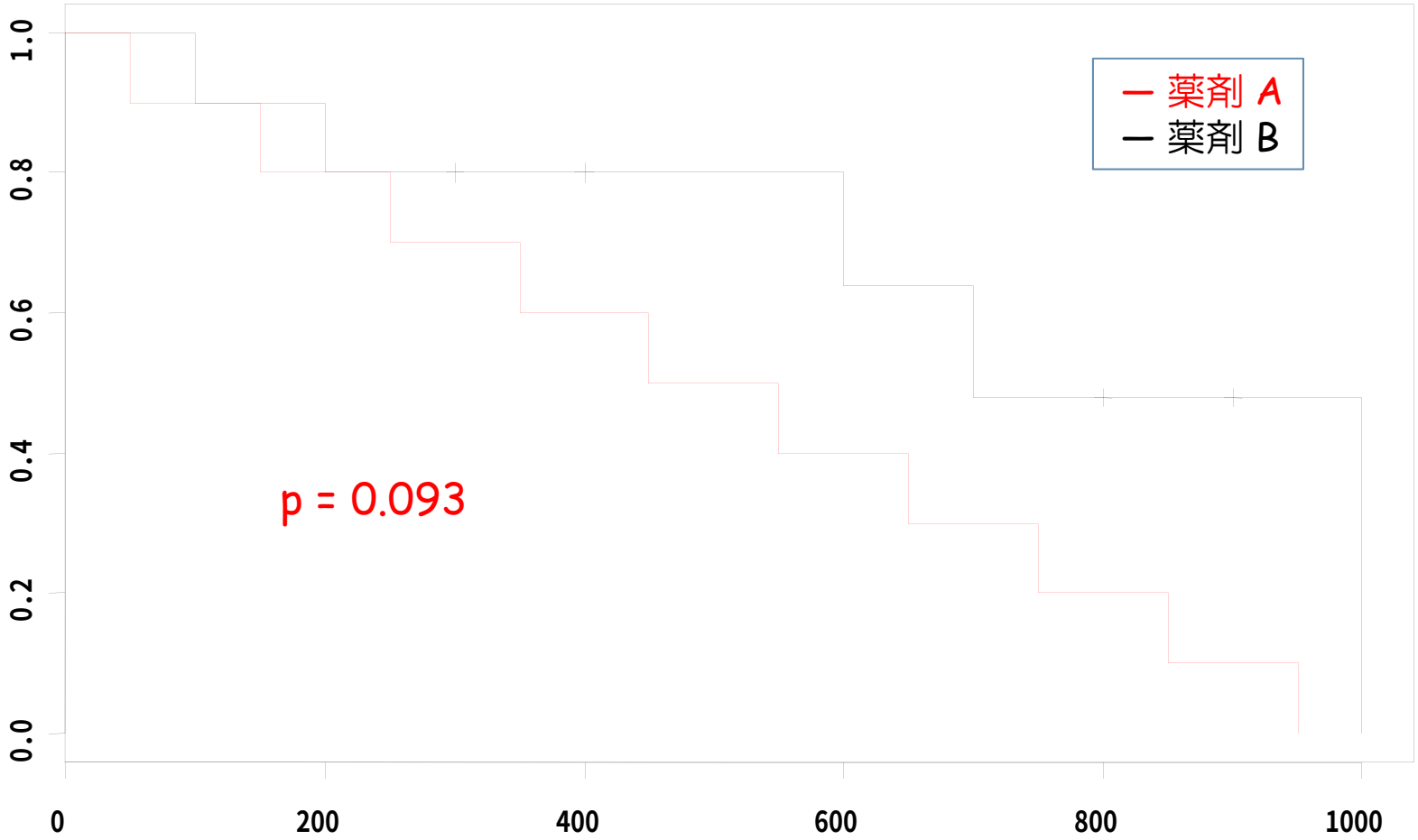
- 群間比較を行う場合 ⇒ 直線が重なっているかどうかを見る
  - 直線が重なっている　：生存関数が同じ　（ログランク検定で p 値が大きい状態）
  - 直線が重なっていない：生存関数が異なる（ログランク検定で p 値が小さい状態）

カプラン・マイヤープロットとログランク検定

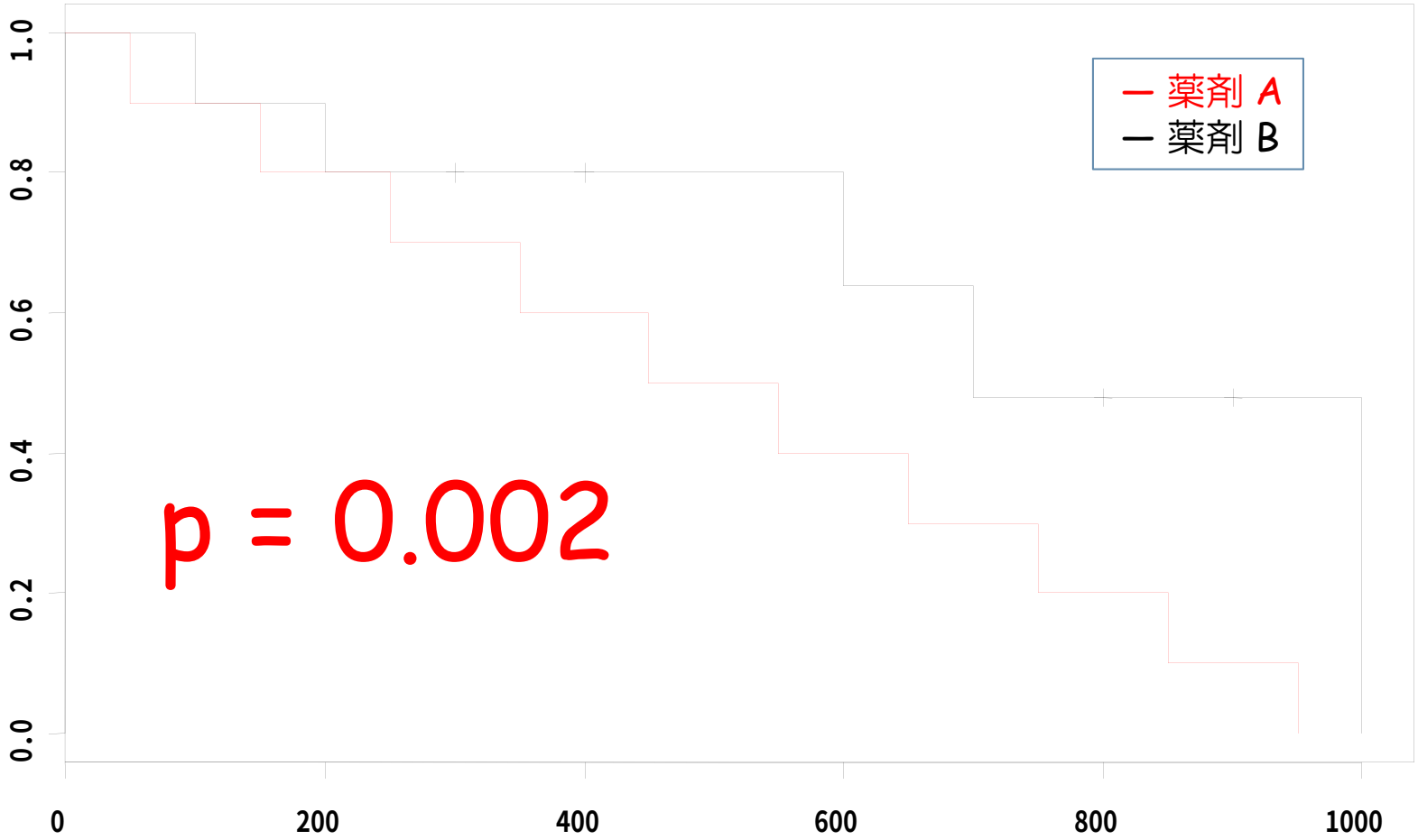


- 群間比較を行う場合 ⇒ 直線が重なっているかどうかを見る
  - 直線が重なっている　：生存関数が同じ　（ログランク検定で p 値が大きい状態）
  - 直線が重なっていない：生存関数が異なる（ログランク検定で p 値が小さい状態）

【参考】検出力はイベント数に依存する



（例数と）イベント数を 3 倍すると曲線の形は変わらないが…



※ 生存時間解析の検出力はイベント数（≠ 患者さんの数）に依存する
※ いくら患者さんの数が多くても，イベントが発現しないと検出力は上がらない

イベントが起こるまでの時間に関するログランク検定

- 「薬剤 A の生存関数（イベントが起こるまでの時間）」と「薬剤 B の生存関数」が 等しいかどうかを検定する
  - p = 4.9%，有意水準 5% で検定すると結果は有意
  - 有意なので生存関数（イベントが起こるまでの時間）は等しくない

```
> survdiff(Surv(DAY,Y) ~ GROUP, data=AB)                    # ログランク検定

Call:
survdiff(formula = Surv(DAY, Y) ~ GROUP, data = AB)

          N Observed Expected (O-E)^2/E (O-E)^2/V
GROUP=A 20       12     8.04      1.95      3.85
GROUP=B 20        5     8.96      1.75      3.85

 Chisq=3.9  on 1 degrees of freedom, p=0.0497 ← 検定結果（ p 値 = 約 4.9%）
```

イベントが起こるまでの時間に関するログランク検定

1. 比較の枠組み ⇒薬剤 *A* と薬剤 B の生存関数を比較する
2. 比較するものの間に差がないという仮説（帰無仮説 H_0）を立てる
   ⇒ 帰無仮説 H_0 ：薬剤 *A* の生存関数 = 薬剤 B の生存関数
3. 帰無仮説とは裏返しの仮説（対立仮説 H_1）を立てる
   ⇒ 対立仮説 H_1 ：薬剤 *A* の生存関数 ≠ 薬剤 B の生存関数
4. 帰無仮説が成り立つという条件の下で，手元にあるデータ（よりも極端なこと）が起こる確率（= p 値）を計算 ⇒ p = 0.0497（4.9%）
5. 対立仮説 H_1 が正しいと結論 ⇒「生存関数は異なる」と解釈する
6. 「生存関数は異なる」 &「薬剤 *A* の曲線の方が薬剤 B よりも下」の合わせ技で「薬剤 *A* の無発生割合 < 薬剤 B の無発生割合」，つまり「薬剤 *A* の累積発生割合 > 薬剤 B の累積発生割合」と結論付ける

本日のメニュー

1. イントロ
2. イベントの無発生割合と累積発生割合の算出
3. 「イベントが起こるまでの時間」の比較
4. その他
   - ハザードについて
   - 人年法によるハザードの計算

「率」とは

- ある事象が単位時間（例えば 1 年）の間に起こった数
  ⇒ 時速（ 1 時間あたりに走る距離）のようなものとイメージ出来る
- 「200人を 1 年間観察した結果，6 人が死亡した」場合：
  「死亡数（ 6 人）」を「のべ観察時間（200人× 1 年＝200人年）」で割り算した値 ＝ 0.03（人/年）が率
  ⇒「 1 人年あたり 0.03 人が死亡する」と解釈する
- 上記の「0.03（人/年） 」に 1000 をかけて
  「 1000人年あたり 30 人が死亡する
  ⇒「1000人を 1 年間観察した場合，6人が死亡する」とも解釈出来る

※「割合」は「時間」の概念がないものに関する指標であるのに対し，「率」は単位時間の間に起こった頻度，と「時間」の概念が入った指標となっているのが相違点

生存関数とハザードと比例ハザード性

- ある時点 t におけるイベントの無発生割合を，生存時間解析では「生存関数」と呼び，関数 S(t) で表す
- カプランマイヤー法で求めたイベントの無発生割合（生存関数 S(t) ）の算出方法とは別に「ある時点 t における 瞬間的な イベント発生率」を表す「ハザード」という概念がある
  - 「瞬間イベント発生率」であるハザードは，例えばイベントが「死亡」であれば「（瞬間）死亡率」と解釈することが出来る
- よくやる解析方法として，注目する 2 つの群（例えば薬剤 A と薬剤 B ）のハザードの比（ハザード比）を計算する方法がある
  - 多くの場合，ハザード比をリスク比のように解釈することが出来る
  - ハザードは 0 から ∞ の値をとるもので，ハザード関数 h(t) で表す
  - 数学的には，時間 t の直前まで生存した人が次の瞬間に死亡する条件付き確率

【参考】各種関数

- T：生存時間を表す確率変数とする

$$S(t) = P(T \geq t) \qquad \text{(生存関数)} \leftarrow \text{無発生率}$$

$$F(t) = 1 - S(t) \qquad \text{(累積分布関数)} \leftarrow \text{発生率}$$

$$h(t) = \frac{f(t)}{S(t)} \qquad \text{(ハザード関数)} \leftarrow \text{瞬間発生率}$$

$$H(t) = \int_0^t h(u)du = -\log S(t) \qquad \text{(累積ハザード関数)}$$

例 T が指数分布 $f(t) = \lambda \exp(-\lambda t)$ に従うとき，

$$F(t) = 1 - \exp(-\lambda t) \qquad \text{(累積分布関数)}$$

$$S(t) = 1 - S(t) = \exp(-\lambda t) \qquad \text{(生存関数)}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda \qquad \text{(ハザード関数)}$$

$$H(t) = -\log S(t) = \lambda t \qquad \text{(累積ハザード関数)}$$

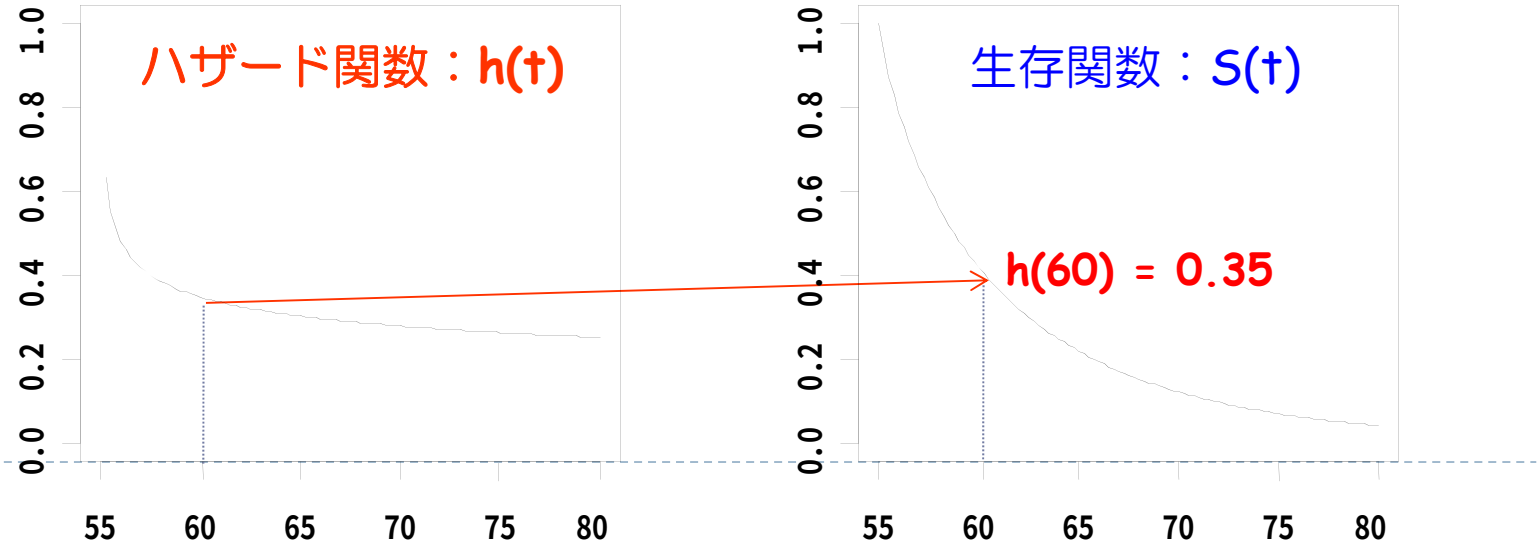
※ $h(t) = -\frac{d}{dt}\log S(t)$ という関係もあり

ハザード関数 ＝ 瞬間イベント発症率

- イベントが死亡の場合
  - 瞬間死亡率
  - 時間 t の直前まで生存した人が次のΔ t の期間に死亡する条件付き確率

例：60 歳の人の死亡率の例

- 60 歳で死亡するためには 60 歳まで生存する必要がある
- 「60 歳まで生存した」という条件の下で、死亡する条件付き確率

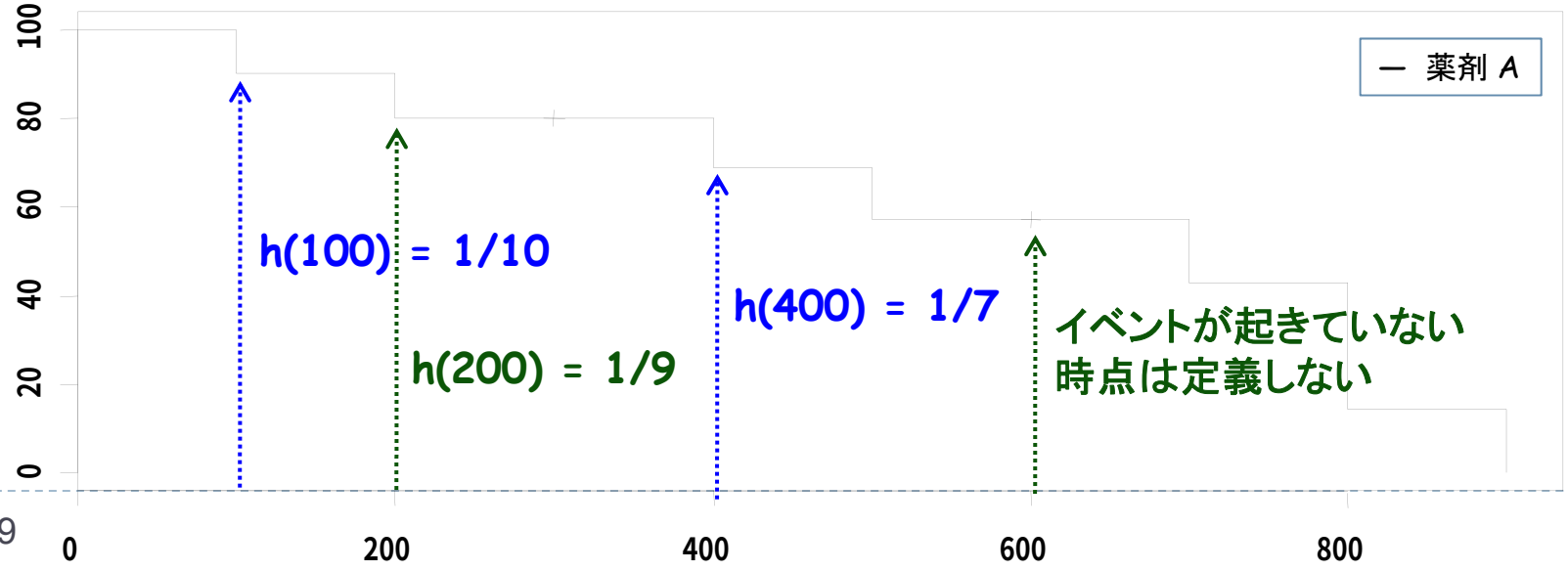


カプラン・マイヤー法で推定した場合のハザード

薬剤	イベント	時間（日）
A	あり	100
A	あり	200
A	なし	300
A	あり	400
A	あり	500
A	なし	500
A	あり	700
A	あり	800
A	あり	800
A	あり	900

- イベントが起きた時点でのみ計算可
  ⇒ それ以外の時点は定義できない… が，
  平滑化して h(t) の関数を求める方法あり
- イベント数が多くなければ安定しない
- もし，薬剤間のハザード比を求める場合
  時点ごとでバラバラになる可能性大

ハザード関数の Nelson-Aalen 推定量：$h(t_i) = d_i/n_i$（d_i：イベント数，n_i：リスク集合）



生存関数とハザードと比例ハザード性

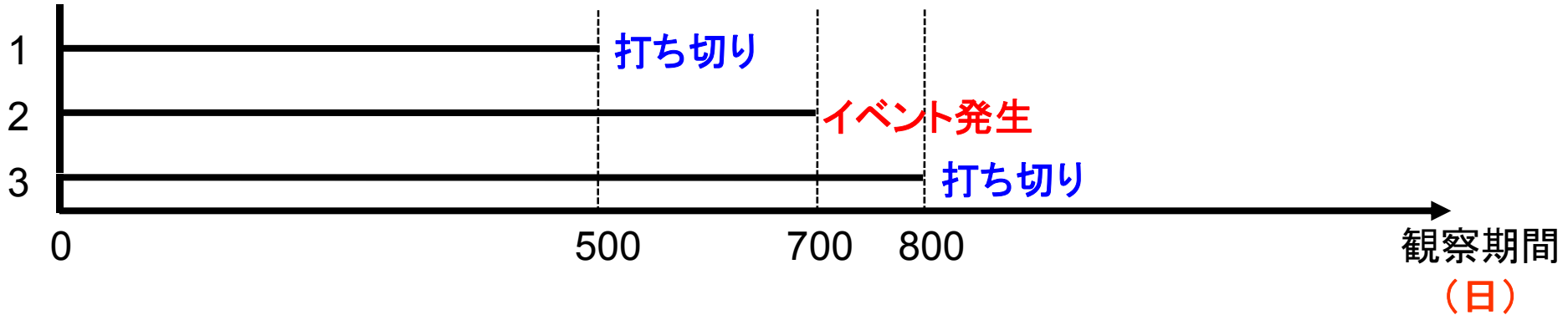
- 「ハザード」は「瞬間イベント発生率」なので，時間ごとにコロコロ変わってしまい解釈しにくいが，次回紹介する「Cox 回帰分析」では「注目する 2 群のハザード比がどの時点でも一定となる」，すなわち「比例ハザード性」を仮定して解析を行う

例：薬剤 B に対する薬剤 A のハザード比（　リスク比の生存時間解析版）が 1.5 であったとき，もし比例ハザード性が成り立っている場合は

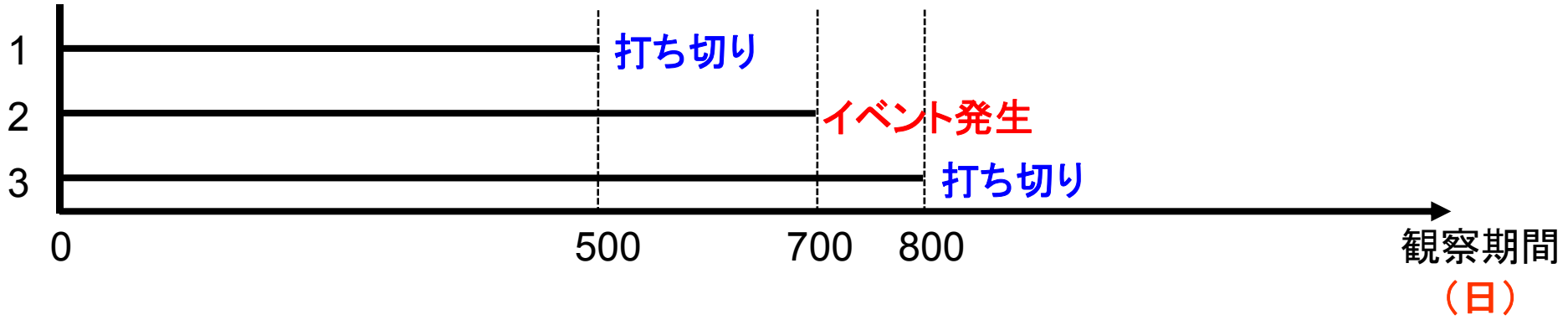
- 観察開始日から 1 日後のハザード比　　＝ 1.5
- 観察開始日から 200日後のハザード比　＝ 1.5
- 観察開始日から 30000日後のハザード比 ＝ 1.5　　ということになる

人年法によるハザードの計算

- 前頁では -logS(t) を t で微分してハザード（瞬間イベント発生率）を求める方法を紹介した
- 今度は「人年法」という方法によりハザード（イベント発生率）を計算してみる
- 例として，うつ病を患っている 3 人の患者さんのデータを使う



人年法によるハザードの計算



- 人年法によるイベント発生率

  ＝ イベント発生数÷総観察時間 ＝ 1 ÷（500+700+800）＝ 0.0005 (/人日)

  ＝ イベント発生数÷(総観察時間÷365.25) ＝ 0.0005×365.25 ＝ 0.1826(/人年)

- イベント発生率を 2 つの別の単位で求めてみたが，結果の解釈は・・・
  - イベント発生率（日）：1人を 1 日観察したときにイベントが発生する率
  - イベント発生率（年）：1人を 1 年観察したときにイベントが発生する率
  - 「人年法」により算出したイベント発生率（年）は「1 人年あたりのイベント発生率は 0.1826（人/年）である」という風に表現する

本日のメニュー

1. イントロ
2. イベントの無発生割合と累積発生割合の算出
3. 「イベントが起こるまでの時間」の比較
4. その他
   - ハザードについて
   - 人年法によるハザードの計算

参考文献

- 統計学（白旗 慎吾 著，ミネルヴァ書房）
- ロスマンの疫学（Kenneth J. Rothman 著，矢野 栄二 他翻訳，篠原出版新社）
- Applied Survival Analysis （Hosmer & Lemeshow，Wiley）
- The R Tips 第 2 版（オーム社）
- R 流！イメージで理解する統計処理入門（カットシステム）

Rで統計解析入門

終