Analiza przeżycia

Piotr Dzierza

Uniwersytet Wrocławski

11.03.2020

Spis treści

Wstęp

Analiza przeżycia jest jednym z działów statystyki matematycznej, w której głównym celem jest badanie nieujemnej zmiennej losowej opisującej czas do pewnego określonego zdarzenia. Zdarzenia te są inne dla różnych dziedzin nauki. W medycynie często interesujący jest czas nawrotu choroby czy śmierci pacjenta. Dla przemysłu ważnym może być czas do awarii sprzętu. Ekonomistów ciekawić może okres, który mija nim osoba bezrobotna znajdzie zatrudnienie.

Wstęp

- Specjalny typ danych ciągłych - czas do określonego zdarzenia
- Nazewnictwo:
  1. Analiza przeżycia (medycyna)
  2. Analiza niezawodności (przemysł)
  3. Analiza historii zdarzeń (demografia)
  4. Analiza trwania (ekonomia i nauki społeczne)
- Dodatkowa własność danych - cenzura obserwacji

Podstawowe założenia

Będziemy zakładać, że obserwujemy zmienną losową X o rozkładzie z dystrybuantą F taką, że

- $F(0) = 0$,
- F jest ciągła i różniczkowalna oraz $F'(t) = f(t)$.

Funkcje przeżycia

Funkcję o równaniu

$$S(t) = 1 - F(t), \quad t \geq 0$$

nazywać będziemy funkcją przeżycia. Opisuje ona prawdopodobieństwo, że zdarzenie nie zajdzie do czasu t.

Funkcja hazardu

Funkcją hazardu (intensywności awarii) nazywamy funkcję

$$h(t) = \lim_{\triangle t \to 0} \frac{\mathbb{P}(X \leq t + \triangle t | X > t)}{\triangle t}.$$

Można ją interpretować jako prawdopodobieństwo zaobserwowania pewnego zdarzenia w chwili $\triangle t$ jeśli nie zostało ono zaobserwowane do chwili t.

Funkcja skumulowanego hazardu

Funkcją skumulowanego hazardu nazywamy funkcję

$$H(t) = \int_0^t h(u)du.$$

Przy przyjętych założeniach $S(t) = exp(-H(t))$.

Przykłady funkcji hazardu

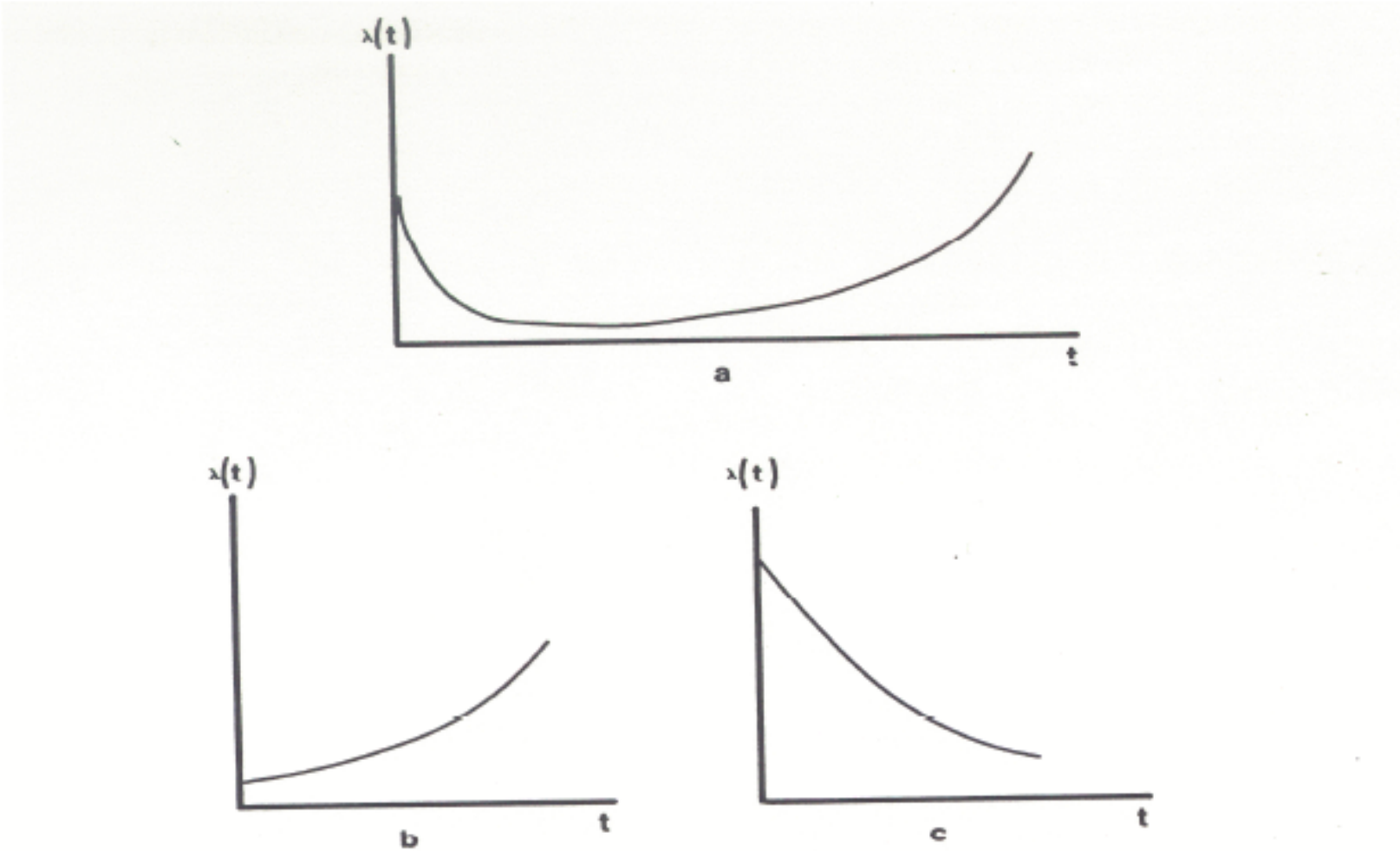


Figure 1.1 Some types of hazard functions: (a) hazard for human mortality; (b) positive aging; (c) negative aging.

Cenzura danych

Jednym z głównych problemów w analizie przeżycia jest cenzura części danych. Dane cenzurowane najczęściej otrzymuje się w wyniku różnych eksperymentów. Cenzurę rozumie się tutaj jako brak dostępności do niektórych informacji, który spowodowany był na przykład zaniechaniem dalszych badań.

Główne typy cenzurowania

Oznaczmy przez T czas zdarzenia.

- Cenzurowanie prawostronne - obserwujemy tylko dolną granicę czasu zdarzenia ($T > C$).
- Cenzurowanie lewostronne - obserwujemy tylko górną granicę czasu zdarzenia ($T < C$).
- Cenzurowanie przedziałowe - obserwujemy dwa punkty czasowe, między którymi było zdarzenie.

Główne typy cenzurowania

- Dane cenzurowane I-go typu - ustalamy maksymalny czas trwania t_0 badania. Wynikiem obserwacji jest wektor $(t_1, \ldots, t_n)$, gdzie t_i jest realizacją zmiennej losowej $T_i = min\{X_i, t_0\}$, X_i to niezależne czasy zdarzeń o jednakowym rozkładzie.
- Dane cenzurowane II-go typu - eksperyment z którego pochodzą dane przeprowadza się do wystąpienia pewnej liczby k zdarzeń, reszta obserwacji bez zdarzeń zostaje ocenzurowana przez wartość k-tego zdarzenia.

Cenzurowanie nieinformatywne

Niech $X_1, \ldots, X_n$ będą niezależnymi zmiennymi losowymi opisującymi czasy życia. Niech $C_1, \ldots, C_n$ będą niezależnymi zmiennymi losowymi, że $X_1, \ldots, X_n, C_1, \ldots, C_n$ są wzajemnie niezależne. Obserwujemy wtedy pary zmiennych (T_i, δ_i), gdzie

$$T_i = min\{X_i, C_i\}, \quad \delta_i = \mathbf{1}(X_i \leq C_i).$$

Takie cenzurowanie nazywamy nieinformatywnym.

Mechanizm prawostronnego cenzurowania

Trzema typowymi sposobami prawego cenzurowania są:

- ucięcie administracyjne związane z końcem badania,
- wycofanie się pacjenta z dalszych badań,
- zerwanie kontaktu z badaczami przez pacjenta.

Fundamentalne założenie: cenzurowanie jest niezależne i nieinformatywne.

Cenzurowanie prawostronne - przykład

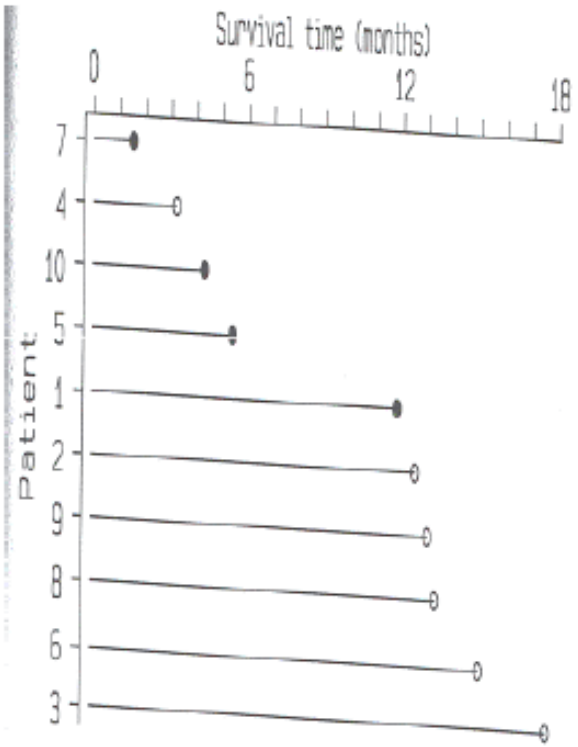


Figure 13.2 Figure 13.1 reorganized to correspond to method of analysis.

Pat #	Survival time (mths.)	Death indicator (1=yes)
k	t_k	d_k
1	11.8	1
2	12.5	0
3	17.6	0
4	3.2	0
5	5.4	1
6	15.0	0
7	1.5	1
8	13.3	0
9	13.0	0
10	4.3	1

Cenzurowanie prawostronne - przykład

Załóżmy, że interesuje nas informacja o prawdopodobieństwie, że pacjent nie przeżyje 6 miesięcy. W zależności od tego jak potraktujemy pacjenta numer 4, możemy rozważyć kilka pomysłów estymacji:

- prawdopodobieństwo wyniesie $\frac{4}{10}$, gdy przyjmiemy, że ocenzurowany pacjent przeżył 6 miesięcy,
- prawdopodobieństwo wyniesie $\frac{3}{10}$, gdy przyjmiemy, że ocenzurowany pacjent nie przeżył 6 miesięcy,
- prawdopodobieństwo wyniesie $\frac{3}{9}$, gdy odrzucimy pacjenta numer 4 z danych.

Pierwszy przypadek powoduje ryzyko przeszacowania, drugi ryzyko niedoszacowania, a trzeci wiąże się z utratą części informacji.

Estymator Kaplana-Meiera

Estymator Kaplana-Meiera funkcji przeżycia opiera się na pomyśle, że aby przeżyć t-ty punkt czasowy, trzeba najpierw przetrwać $t-1$ wcześniejszych punktów.
Przy założeniu, że $S(0)=1$, powyższe można zapisać jako

$$S(t) = S(t-1) \cdot \mathbb{P}(\text{przetrwanie punktu czasowego } t).$$

Estymator Kaplana-Meiera

Załóżmy, że obserwujemy d różnych czasów zdarzeń i cenzury t_j, które są uporządkowane rosnąco,

$$t_1 < t_2 < \ldots < t_d.$$

Przez n_j oznaczmy moc zbioru ryzyka w czasie t_j, czyli liczbę obserwowanych osobników tuż przed momentem t_j. Przez d_j oznaczmy liczbę zdarzeń w czasie t_j. Estymator K-M funkcji przeżycia wyraża się wzorem

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right).$$

Przykład 1 - choroba lokomocyjna

Posiadamy dane z eksperymentu badawczego, który zakładał sprawdzenie wytrzymałości uczestników na turbulencje o częstotliwości $0,167Hz$ oraz przyspieszeniu $0,111G$. Interesującym zdarzeniem był moment pierwszych wymiotów. W tym dwugodzinnym badaniu brało udział 21 osób, z czego tylko dwóch uczestników przerwało badania "na żądanie", co jest równoznaczne z prawostronną cenzurą. Dodatkowo, 14 osób ukończyło badania bez zdarzenia.

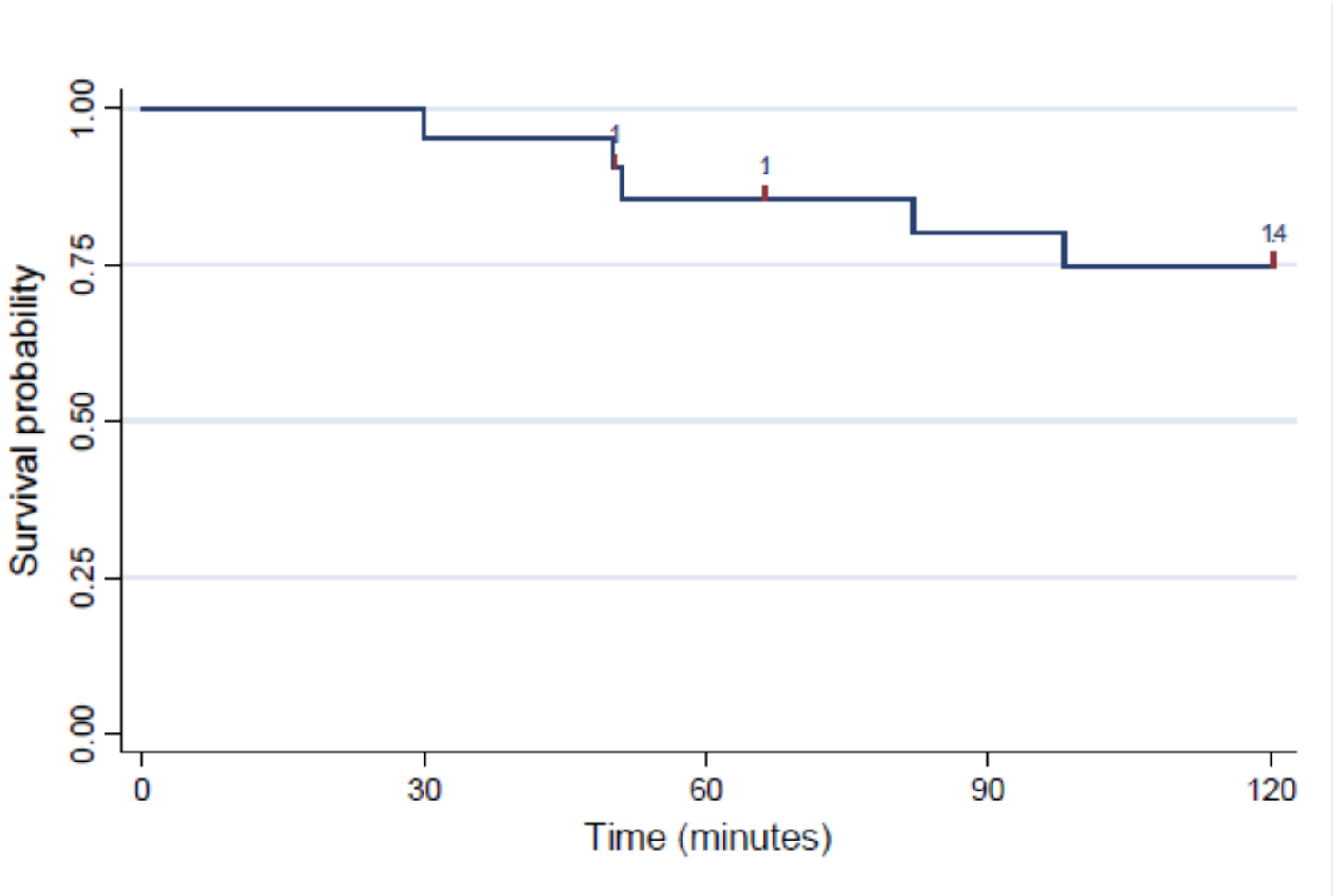
Przykład 1 - choroba lokomocyjna

Subject	Time (minutes)	Event (1=yes)
1	30	1
2	50	1
3	50	0
4	51	1
5	66	0
6	82	1
7	92	1
8	120	0
...	...	...
21	120	0

Przykład 1 - obliczenia

Time t	Risk-set at t R_t	Events at t D_t	Survival (0 events) of t $P_t=(R_t-D_t)/R_t$	Survival function (0 events in 0-t) $S(t)=S(t-1)\cdot P_t$
1	21	0	21/21	1·21/21=1
2	21	0	21/21	1·21/21=1
...	21	0	21/21	1·21/21=1
29	21	0	21/21	1·21/21=1
30	21	1	20/21	1·20/21=0.952
31	20	0	20/20	0.952·20/20=0.952
...	20	0	20/20	0.952·20/20=0.952
49	20	0	20/20	0.952·20/20=0.952
50	20	1	19/20	0.952·19/20=0.905
51	18	1	17/18	0.905·17/18=0.854
...	17	0	17/17	0.854·17/17=0.854
66	17	0	17/17	0.854·17/17=0.854
67	16	0	16/16	0.854·16/16=0.854
...	16	0	16/16	0.854·16/16=0.854
82	16	1	15/16	0.854·15/16=0.801
...	15	0	15/15	0.801·15/15=0.801
92	15	1	14/15	0.801·14/15=0.748
...	14	0	14/14	0.748·14/14=0.748
120	14	0	14/14	0.748·14/14=0.748

Przykład 1 - funkcja przeżycia



Przykład 1 - wnioski

- $S(t)$ zmienia się tylko w czasach zdarzenia.
- Obserwacje cenzurowane zmieniają moc zbioru ryzyka.
- Dla tych danych $S(t)$ jest niezdefiniowane dla $t > 120$.

Przykład 2

Przyjrzyjmy się ponownie danym z przykładu o cenzurowaniu prawostronnym.

Pat #	Survival time (mths.)	Death indicator (1=yes)
k	t_k	d_k
1	11.8	1
2	12.5	0
3	17.6	0
4	3.2	0
5	5.4	1
6	15.0	0
7	1.5	1
8	13.3	0
9	13.0	0
10	4.3	1

Przykład 2

Time (mths.)	Death (1=yes)	S(t)
1.5	1	90.0%
3.2	0	90.0%
4.3	1	78.7%
5.4	1	67.5%
11.8	1	56.2%
12.5	0	56.2%
13.0	0	56.2%
13.3	0	56.2%
15.0	0	56.2%
17.6	0	56.2%

Przykład 2

Mieliśmy wtedy:

- prawdopodobieństwo wyniesie $\frac{4}{10} = 40\%$, gdy przyjmiemy, że ocenzurowany pacjent przeżył 6 miesięcy,
- prawdopodobieństwo wyniesie $\frac{3}{10} = 30\%$, gdy przyjmiemy, że ocenzurowany pacjent nie przeżył 6 miesięcy,
- prawdopodobieństwo wyniesie $\frac{3}{9} = 33,3\%$, gdy odrzucimy pacjenta numer 4 z danych.

Estymator Kaplana-Meiera wynosi w tym przypadku $67,5\%$, czyli prawdopodobieństwo przeżycia wynosi $32,5\%$.

Mediana czasu do zdarzenia

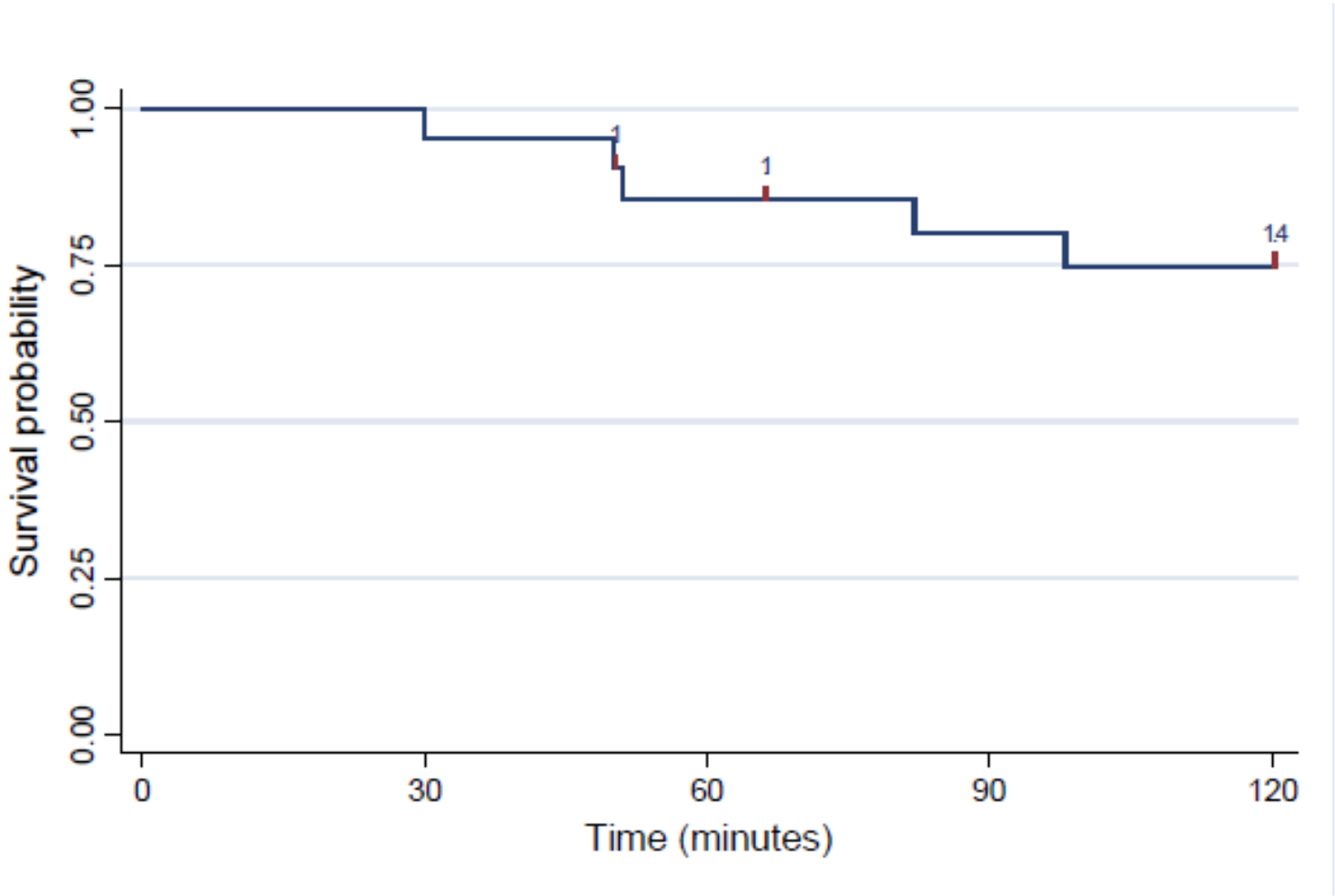
Interesuje nas moment w czasie t_{med}, dla którego

$$S(t_{med}) = 50\%.$$

Estymatorem będzie najmniejszy czas zdarzenia t_k dla którego $S(t_k) < 50\%$ lub w przypadku, gdy $S(t) = 50\%$ zachodzi dla pewnego przedziału, bierzemy środek tego przedziału.
Problemy:

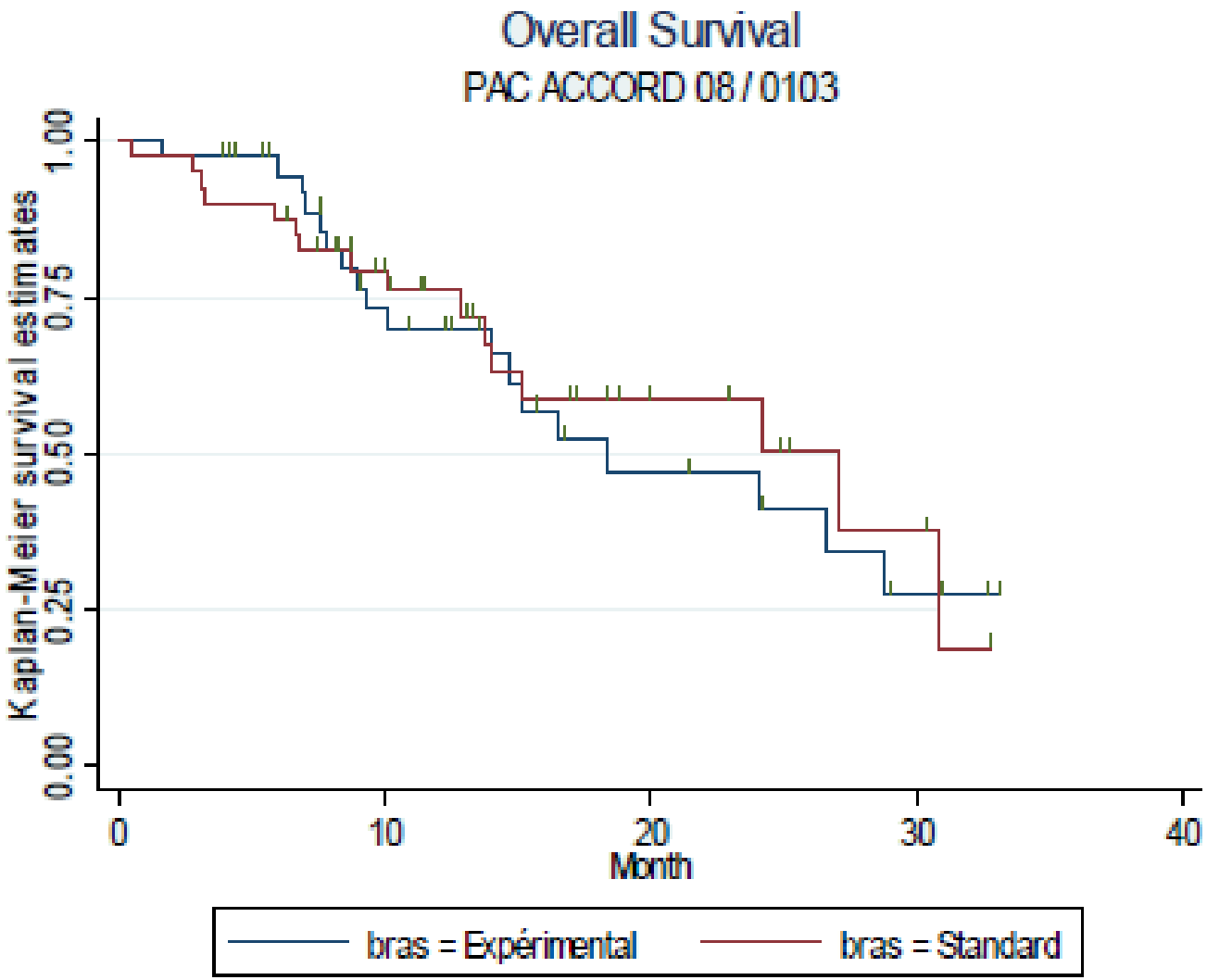
- estymator może nie istnieć,
- ciężko uzyskać przedział ufności,
- estymator może mieć dużą wariancję.

Mediana czasu do zdarzenia



W tym przypadku estymator mediany czasu do zdarzenia nie istnieje.

Mediana czasu do zdarzenia



Mediana czasu do zdarzenia

Estymatory mediany czasu do zdarzenia:

- dla *bras* = *Experimental* wynosi 18 miesięcy,
- dla *bras* = *Standard* wynosi 27 miesięcy.

Zamienienie jednej obserwacji w grupie *bras* = *Standard* z cenzurowania na zdarzenie mocno zmniejszy estymator mediany (pionowe kreseczki na wykresach oznaczają obecność obserwacji cenzurowanej).