

# Raport Lista 2

## Analiza Przeżycia

Dominik Kowalczyk i Matylda Mordal

r Sys.Date()

### Spis treści

<b>1</b>	<b>Lista 5</b>	<b>2</b>
1.1	Zadanie 1 . . . . .	2
1.2	Zadanie 2 . . . . .	4
1.3	Zadanie 3 . . . . .	5
<b>2</b>	<b>Lista 6</b>	<b>10</b>
2.1	Zadanie 1 . . . . .	10
2.2	Zadanie 2 . . . . .	12
<b>3</b>	<b>Lista 7</b>	<b>14</b>
3.1	Zadanie 1 . . . . .	14
3.2	Zadanie 2 . . . . .	16
<b>4</b>	<b>Lista 8</b>	<b>19</b>
4.1	Zadanie 1 . . . . .	19
4.2	Zadanie 2 . . . . .	21

# 1 Lista 5

## 1.1 Zadanie 1

W zadaniu analizujemy przebieg funkcji przeżycia czasu do remisji choroby u pacjentów leczonych dwoma różnymi lekami, oznaczonymi jako lek A oraz lek B. Celem jest porównanie oszacowań funkcji przeżycia uzyskanych przy użyciu dwóch klasycznych metod estymacji stosowanych w analizie danych cenzurowanych:

W zadaniu analizujemy dane dotyczące czasu do remisji choroby u pacjentów, których leczono dwoma różnymi lekami oznaczonymi jako lek A oraz lek B. Zbiór danych zawiera:

- czasy do remisji,
- wskaźniki cenzurowania,
- informację o grupie: lek A lub lek B.

Dane mają charakter prawostronnie cenzurowany.

Celem jest porównanie oszacowań funkcji przeżycia uzyskanych przy użyciu dwóch metod estymacji:

- Estymator Kaplana-Meiera

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \frac{r_i - d_i}{r_i}, \quad \text{gd}y \ 0 \leq t \leq t_{(n)},$$

- Estymator Fleminga-Harringtona

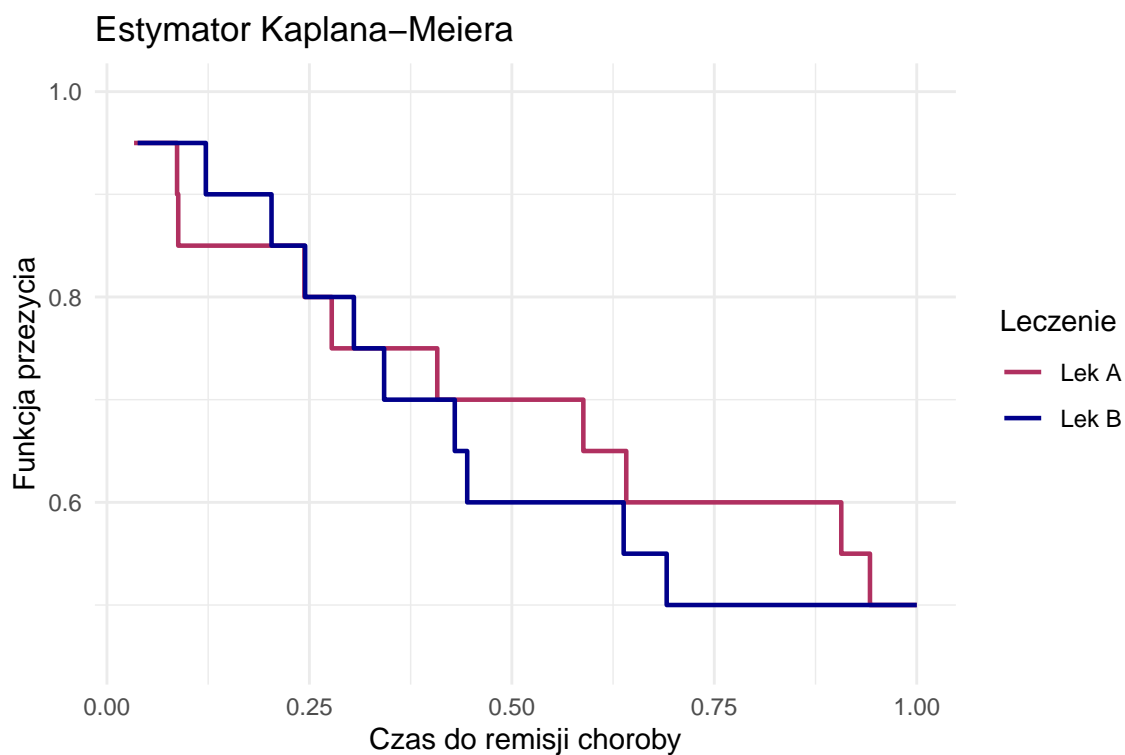
$$\tilde{S}(t) = \exp \left( - \sum_{j:t_{(j)} \leq t} \frac{d_j}{r_j} \right), \quad \text{gd}y \ 0 \leq t \leq t_{(n)},$$

gdzie:

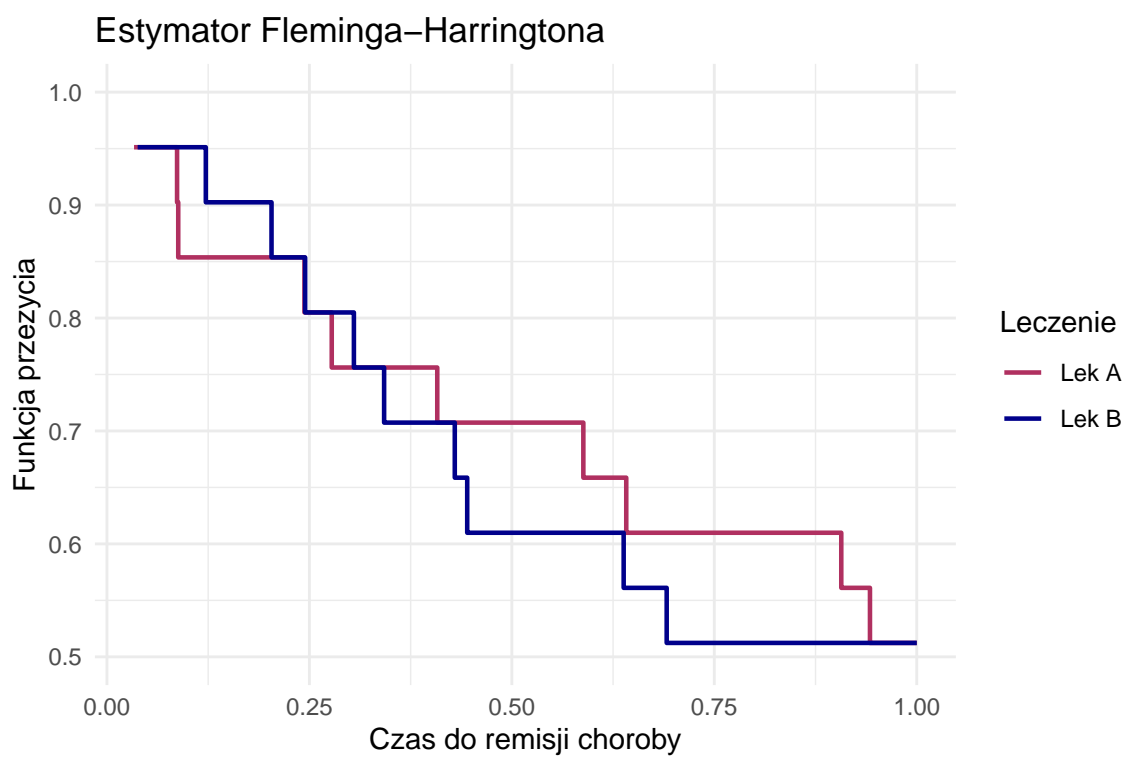
- $r_i = \sum_{j=1}^n \mathbf{1}_{[t_{(i)}, \infty)}(T_j)$  - liczba jednostek będących w stanie ryzyka w punkcie  $t_{(i)}$ ,
- $d_i = \sum_{j=1}^n \mathbf{1}_{\{t_{(i)}\}}(t_j) \mathbf{1}_{\{1\}}(\delta_j)$  - liczba zdarzeń, które zaszły w chwili  $t_{(i)}$  (nieocenzurowane).

### Kod

Zastosowana funkcja *survfit* wyznacza estymatory funkcji przeżycia osobno dla pacjentów leczonych lekiem A oraz lekiem B. Estymator Kaplana-Meiera wywoływany jest za pomocą funkcji *survfit(Surv(Czas, Cenzura) ~ 1, data = ...)*. Pole *data = Dane\_A* lub *data = Dane\_B* wskazuje odpowiedni podzbiór danych dla danej grupy. Analogicznie, dodając argument *type = "fleming-harrington"*, otrzymujemy estymatory Fleminga-Harringtona.



Rysunek 1: Estymatory Kaplana-Meiera funkcji przeżycia dla Leku A i Leku B



Rysunek 2: Estymatory Fleminga-Harringtona funkcji przeżycia dla Leku A i Leku B

## Interpretacja wykresów

Na rysunku 1 i 2 przedstawiono oszacowania funkcji przeżycia dla pacjentów leczonych lekiem A i lekiem B, odpowiednio przy użyciu estymatora Kaplana-Meiera oraz estymatora Fleminga-Harringtona. Można zauważyć, że estymator Fleminga-Harringtona daje krzywe położone nieco wyżej niż estymator Kaplana-Meiera, szczególnie w późniejszych momentach obserwacji. Wynika to z faktu, że opiera się on na estymatorze Nelsona-Aalena, który zazwyczaj prowadzi do łagodniejszego spadku funkcji przeżycia, a tym samym do nieco wyższych wartości.

Obie funkcje przeżycia mają zbliżony przebieg, choć widoczne są subtelne różnice. Krzywa dla leku B początkowo utrzymuje się nieco wyżej niż dla leku A, co wskazuje na większe szanse braku remisji w pierwszych etapach. W dalszej części obserwacji lek B wykazuje wyraźniejsze spadki, co może sugerować szybsze osiągnięcie remisji przez pacjentów z tej grupy.

## 1.2 Zadanie 2

W przypadku gdy  $\delta_{(n)} = 0$ , estymator Kaplana-Meiera  $\hat{S}(t)$  nie jest określony dla  $t > t_{(n)} =: t^+$ . Rozwiązanie Browna, Hollandera i Kowara zakłada oszacowanie „ogona” funkcji przeżycia dla  $t > t^+$  przez odpowiednio dobraną funkcję przeżycia rozkładu wykładniczego.

Odpowiedni parametr  $\vartheta$  rozkładu wykładniczego jest wyznaczany z równania

$$\exp(-\vartheta t^+) = \hat{S}(t^+),$$

i wynosi

$$\vartheta = -\frac{\ln \hat{S}(t^+)}{t^+}.$$

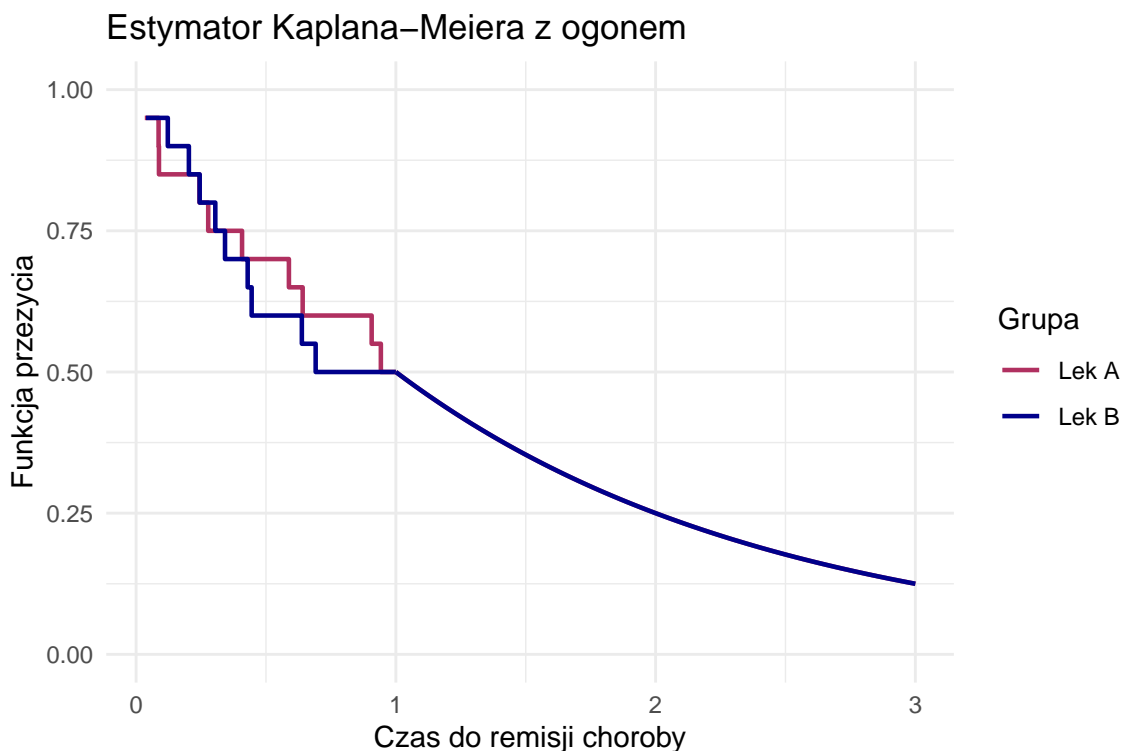
Stąd otrzymujemy, że w przypadku estymatora Browna, Hollandera i Kowara

$$\hat{S}(t) = \exp \left[ \frac{\ln \hat{S}(t^+)}{t^+} t \right], t > t^+.$$

## Kod

Poniżej zaprezentowano implementację funkcji, która szacuje „ogon” estymatora Kaplana-Meiera zgodnie z propozycją Browna, Hollandera i Kowara.

```
# Funkcja ogona
ogon <- function(km) {
  t_plus <- km$time[length(km$time)]
  S_plus <- km$surv[length(km$surv)]
  t_max <- 3
  t_tail <- seq(t_plus, t_max, length.out = 500)
  S_tail <- exp(log(S_plus) / t_plus * (t_tail))
  data.frame(time = t_tail, surv = S_tail)
}
```



Rysunek 3: Estymatory Kaplan-Meiera funkcji przeżycia dla Lek A i Lek B z „ogonem” estymowanym według metody Browna, Hollandera i Kowara

## Interpretacja

Rysunek 3 przedstawia szacowanie funkcji przeżycia czasu do remisji choroby w dwóch grupach pacjentów. Wykres jest podzielony na dwie sekcje: część schodkową, stanowiącą standardowy Estymator Kaplana-Meiera, stanowiącą „ogon” uzyskany metodą Browna, Hollandera i Kowara. Analiza funkcji przeżycia w części schodkowej prowadzi do tych samych wniosków, co w zadaniu 1.1, Lek A początkowo wykazuje szybszą remisję, lecz przez większość czasu obserwacji utrzymuje wyższe wartości funkcji przeżycia niż Lek B.

Ogon estymatora zapewnia ciągłość funkcji przeżycia po  $t^+$ . W części ogona, obie krzywe mają wykładniczy kształt, co pozwala przypuszczać, że w dalszym okresie skuteczność obu leków pozostaje porównywalna.

## 1.3 Zadanie 3

Celem zadania jest zbadanie własności estymatora Kaplana-Meiera z “ogonem” według Browna, Hollandera i Kowara. Generujemy  $M = 1000$  zbiorów danych cenzurowanych I-go typu z uogólnionego rozkładu wykładniczego  $\mathcal{GE}(\lambda, \alpha)$ , przy czym przyjęliśmy parametry  $\alpha = 1.2$  oraz  $\lambda = 0.5$ . Rozważamy różne licznosci próby, odpowiednio  $n = 30, 50, 100$ , a  $t_0$  został wybrany w przybliżeniu jako odwrotność parametru  $\lambda$ , czyli  $t_0 \approx 1/\lambda$ . Na podstawie wygenerowanych danych wyznaczaliśmy oszacowania funkcji przeżycia w punktach  $t_0$  oraz  $2t_0$ , a następnie dla otrzymanych wyników sporządzane są histogramy. Dodatkowo, aby formalnie

ocenić asymptotyczną normalność w punkcie  $t_0$  i  $2t_0$ , zastosowano test Shapiro-Wilka, który jest jednym z najczęściej stosowanych testów normalności.

## Kod

W kodzie przeprowadzono symulację danych w celu zbadania własności estymatora Kaplana-Meiera. Kluczowym elementem jest funkcja *S\_wartosc*, która oblicza wartość funkcji przeżycia w danym punkcie czasu, przy czym dla czasów większych niż ostatnia obserwacja stosowane jest przedłużenie „ogona” zgodnie z propozycją Browna-Hollandera-Kowara. W pętli, dane cenzurowane I-go typu są generowane przy użyciu funkcji *Ityp*. Następnie, dla każdego zbioru danych, wyznaczany jest estymator Kaplana-Meiera za pomocą funkcji *survfit*, a następnie funkcja *S\_wartosc* jest wykorzystywana do wyznaczenia oszacowań  $\hat{S}(t_0)$  oraz  $\hat{S}(2t_0)$ . Ostateczne wyniki są zapisywane w ramce danych *wyniki*.

```
# Parametry
M <- 1000
n_vec <- c(30, 50, 100)
alpha <- 1.2
lambda <- 0.5
t0 <- 1 / lambda

S_wartosc <- function(km, t) {
  t_plus <- km$time[length(km$time)]
  S_plus <- km$surv[length(km$surv)]
  if (t <= t_plus) {
    return(km$surv[max(which(km$time <= t))])
  } else {
    return(exp(log(S_plus) / t_plus * t))
  }
}

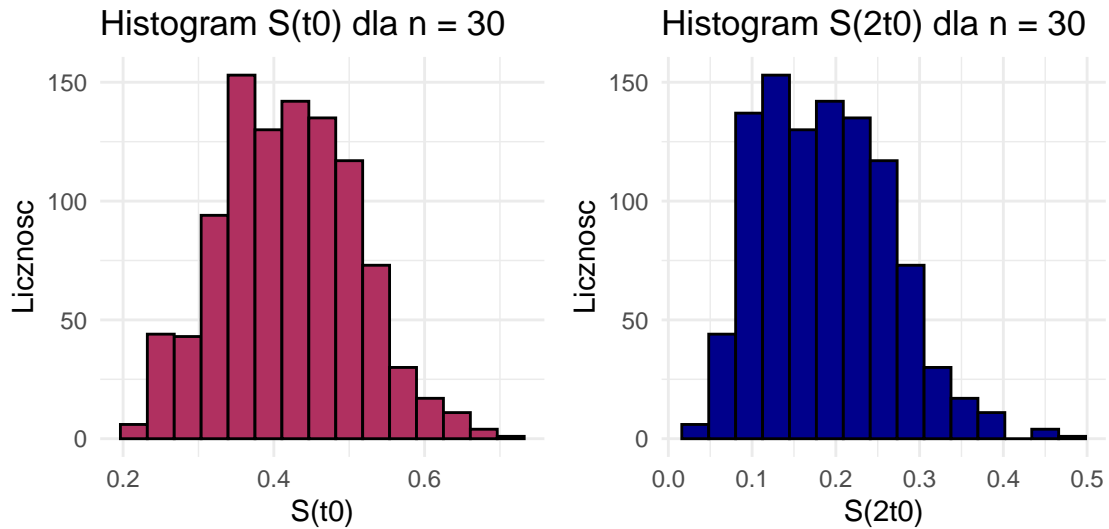
#Symulacja
wyniki <- data.frame()
for (n in n_vec) {
  S_hat_t0 <- numeric(M)
  S_hat_2t0 <- numeric(M)
  for (m in 1:M) {
    #Generujemy dane
    dane <- Ityp(n, alpha, lambda, t0)
    #Tworzymy estymator KM
    km.fit <- survfit(Surv(Czas, Cenzura) ~ 1, data = dane)

    S_hat_t0[m] <- S_wartosc(km.fit, t0)
    S_hat_2t0[m] <- S_wartosc(km.fit, 2*t0)
  }
}
```

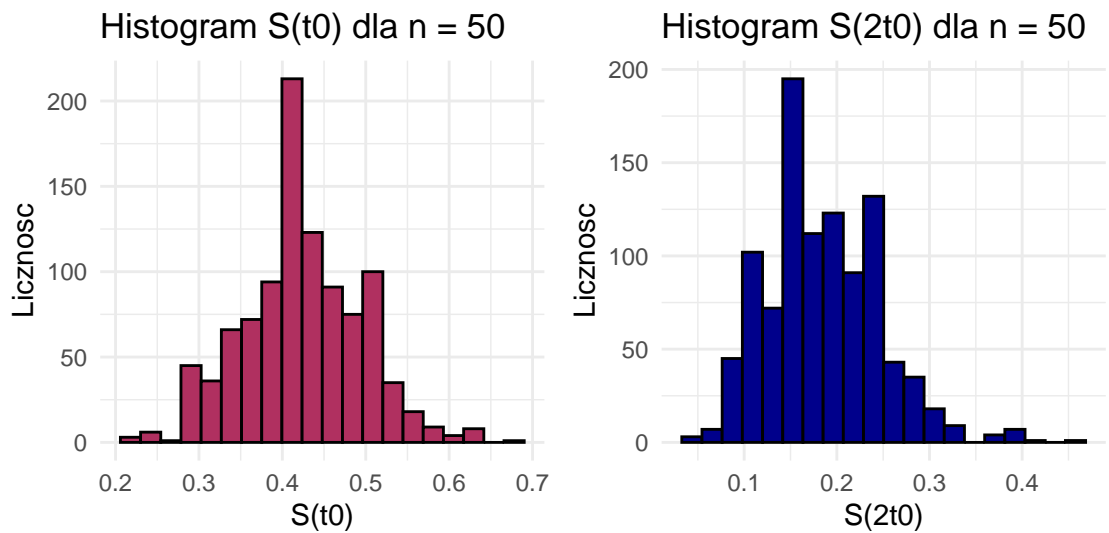
```

dane_n <- data.frame(n = n, S_t0 = S_hat_t0, S_2t0 = S_hat_2t0)
wyniki <- rbind(wyniki, dane_n)
}

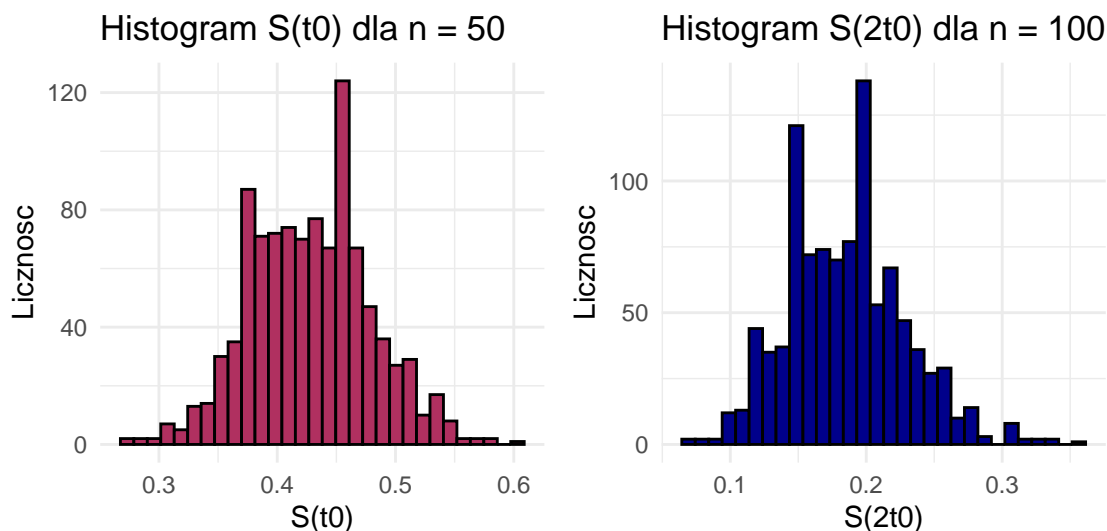
```



Rysunek 4: Histogramy oszacowań estymatora Kaplana-Meiera z „ogonem” według Browna, Hollandera i Kowara funkcji przeżycia w punktach  $t_0$  i  $2t_0$  dla  $n=30$



Rysunek 5: Histogramy oszacowań estymatora Kaplana-Meiera z „ogonem” według Browna, Hollandera i Kowara funkcji przeżycia w punktach  $t_0$  i  $2t_0$  dla  $n=50$



Rysunek 6: Histogramy oszacowań estymatora Kaplana-Meiera z „ogonem” według Browna, Hollandera i Kowara funkcji przeżycia w punktach  $t_0$  i  $2t_0$  dla  $n=100$

Tabela 1: Test Shapiro-Wilka dla estymatora Kaplana-Meiera

n	p-value ( $t_0$ )	p-value ( $2t_0$ )
30	1.19e-08	3.80e-15
50	1.30e-05	2.80e-13
100	7.95e-03	3.32e-08

### 1.3.1 Interpretacja

Na rysunku 4 przedstawiono rozkłady estymatora Kaplana-Meiera w punktach  $t_0$  oraz  $2t_0$  dla stosunkowo małej liczności próby ( $n = 30$ ). Oba histogramy mają niesymetryczny kształt, można zauważyć skośność prawostronną, bardziej widoczną w przypadku  $\hat{S}(2t_0)$ , przez co nie przypomina klasycznego dzwonu charakterystycznego dla rozkładu normalnego. Taki kształt histogramów sugeruje, że przy  $n = 30$  estymator Kaplana-Meiera nie wykazuje asymptotycznej normalności.

Rysunek 5 pokazuje analogiczne rozkłady dla próby  $n = 50$ . W przypadku  $t_0$  rozkład jest bardziej symetryczny niż dla  $n = 30$ , co sugeruje, że wraz ze wzrostem liczności próby estymator jest bardziej skupiony. Dla  $2t_0$  histogram nadal wykazuje skośność i koncentrację przy niższych wartościach, choć minimalnie mniejszą niż w rysunku 4.

Natomiast rysunek 6 przedstawia wyniki dla największej próby  $n = 100$ . Histogram dla  $t_0$  jest już znacznie bardziej zbity i prawie symetryczny. Dla  $2t_0$  nadal wykazuje lekką skośność, choć jest bardziej wygładzony niż w poprzednich przypadkach.

Na podstawie analizy wykresów możemy stwierdzić, że dla punktu  $t_0$  estymator Kaplana-Meiera może być asymptotycznie normalny, ponieważ wraz ze wzrostem liczby prób jego rozkład staje się coraz bardziej symetryczny i zbliża do normalności. Natomiast dla punktu

$2t_0$  jest to trudniejsze do stwierdzenia, ponieważ wykresy wskazują na pewną skośność, a rozkład w tym punkcie jest mniej regularny.

Jednocześnie jednak test Shapiro–Wilka (tabela 1) wykazuje:

- p-value dla  $n=30$ : ekstremalnie małe
- p-value dla  $n=50$ : lekko większe, ale nadal bardzo małe
- p-value dla  $n=100$ : wciąż poniżej poziomu 0.05, jak reszta wyników.

Oznacza to, że formalnie odrzucamy normalność we wszystkich trzech przypadkach, nawet przy  $n=100$ . Wyniki są spójne z histogramami, rozkłady są coraz bardziej regularne, ale nadal nie w pełni normalne.

## 2 Lista 6

### 2.1 Zadanie 1

Celem zadania jest napisanie funkcji do estymacji średniego czasu życia w oparciu o estymator Kaplana-Meiera oraz estymator Fleminga-Harringtona. Obydwa estymatory należy zaimplementować z opcją szacowania ogona (wykładniczego) zaproponowaną przez Browna, Hollandra i Kowara.

#### 2.1.1 Wstęp teoretyczny

Opis oraz wzory dotyczące estymatora Kaplana-Meiera i Fleminga-Harringtona oraz szacowania ogona wykładniczego zostały przedstawione w zadaniu 1.2.

##### *Estymacja punktowa średniej czasu przeżycia*

Wartość oczekiwana czasu życia  $\mu = EX$  dla zmiennej losowej  $X$  (o rozkładzie absolutnie ciągłym względem miary Lebesgue'a) o funkcji przeżycia  $S(t)$  wyraża się wzorem:

$$\mu = \int_0^{\infty} S(t) dt \quad (1)$$

Podstawiając do powyższego wzoru estymator z ogonem wykładniczym możemy otrzymać:

$$\hat{\mu} = \int_0^{t^+} \hat{S}(t) dt + \int_{t^+}^{\infty} [\hat{S}(t^+)]^{t/t^+} dt \quad (2)$$

Drugą całkę możemy przekształcić w postać wykładniczą za pomocą zastosowania  $x = e^{\ln(x)}$  oraz własności logarytmów, wtedy:

$$\hat{\mu} = \int_0^{t^+} \hat{S}(t) dt + \int_{t^+}^{\infty} \exp\left(\frac{\ln(\hat{S}(t^+))}{t^+} \cdot t\right) dt \quad (3)$$

#### 2.1.2 Kod

W pakiecie R implementujemy dwie oddzielne funkcje dla estymatora Kaplana-Meiera oraz Fleminga-Harringtona. Idea polega na tym co pokazaliśmy w wersji teoretycznej - policzeniu dwóch oddzielnych przypadków: wartości pól po wykresie dla przedziału od 0 do  $t^+$  i całki (funkcja - integrate) od  $t^+$  do  $\infty$ , a następnie je zsumować.

```
#Implementacja estymatora Kaplana - Meiera z ogonem
estymator_KM <- function(czas, status){
  KM_fit <- survfit(Surv(czas, status)~1, type="kaplan-meier")
  #wektory czasu i prawd. - musimy dołączyć 0 i 1
  czasy <- c(0, KM_fit$time)
  prawd <- c(1, KM_fit$surv)
  n <- length(czasy)
```

```

mi <- 0

#pola prostokątów (pierwsza całka)
for(i in 1:(n-1)){
  szerokosc <- czasy[i+1] - czasy[i]
  wysokosc <- prawd[i]
  pole_prostokata <- szerokosc * wysokosc
  mi <- mi + pole_prostokata}

t_plus = max(KM_fit$time)
s_plus = min(KM_fit$urv)
#przekształcone S
S_t <- function(t){
  exp(((log(min(KM_fit$urv)))/t_plus)*t)}

#druga całka (ogon)
if(s_plus <= 0){
  ogon <- 0}
else {
  ogon <- integrate(S_t, lower = t_plus, upper = Inf)$value}

wynik_ogon <- mi + ogon
return(c("Bez_ogona" = mi, "Z_ogonem" = wynik_ogon))
}

#Implementacja estymatora Flaminga - Harringtona z ogonem
estymator_FH <- function(czas, status){
  FH_fit <- survfit(Surv(czas, status)~1, type="fleming-harrington")
  #wektory czasu i prawd. - musimy dołączyć 0 i 1
  czasy <- c(0, FH_fit$time)
  prawd <- c(1, FH_fit$urv)
  n <- length(czasy)
  mi <- 0

  #pola prostokątów (pierwsza całka)
  for(i in 1:(n-1)){
    szerokosc <- czasy[i+1] - czasy[i]
    wysokosc <- prawd[i]
    pole_prostokata <- szerokosc * wysokosc
    mi <- mi + pole_prostokata}

  t_plus = max(FH_fit$time)
  s_plus = min(FH_fit$urv)
  #przekształcone S

```

```

S_t <- function(t){
  exp(((log(min(FH_fit$surv)))/t_plus)*t)}

#druga całka (ogon)
if(s_plus <= 0){
  ogon <- 0}
else {
  ogon <- integrate(S_t, lower = t_plus, upper = Inf)$value}

wynik_ogon <- mi + ogon
return(c("Bez_ogona" = mi, "Z_ogonem" = wynik_ogon))
}

```

## 2.2 Zadanie 2

Celem zadania jest oszacowanie średniego czasu życia (przetestowanie kodów z zadania 2.1) dla danych z zadania 3 z listy 2.

### 2.2.1 Kod

```

#dane z zadania 3 listy 2
pA <- c(0.03345514,0.08656403,0.08799947,0.24385821,0.27755032,
        0.40787247,0.58825664,0.64125620,0.90679161,0.94222208)
pB <- c(0.03788958,0.12207257,0.20319983,0.24474299,0.30492413,
        0.34224462,0.42950144,0.44484582,0.63805066,0.69119721)

probka_A <- c(pA, rep(1,10))
cenzura_A <- c(rep(1, 10), rep(0, 10))
probka_B <- c(pB, rep(1,10))
cenzura_B <- c(rep(1, 10), rep(0, 10))

#zastosowanie funkcji z zadania 1 dla danych
srednia_A_KM <- estymator_KM(probka_A, cenzura_A)
srednia_B_KM <- estymator_KM(probka_B, cenzura_B)
srednia_A_FH <- estymator_FH(probka_A, cenzura_A)
srednia_B_FH <- estymator_FH(probka_B, cenzura_B)

```

Tabela 2: Porównanie średniego czasu przeżycia

Lek	Estymator Kaplana-Meiera		Estymator Fleminga-Harringtona	
	Bez ogona	Z ogonem	Bez ogona	Z ogonem
Lek A	0.710791	1.432139	0.717916	1.484004
Lek B	0.672933	1.394281	0.680997	1.447085

### 2.2.2 Interpretacja

Na podstawie porównania danych zamieszczonych w tabeli 2 możemy stwierdzić, że skuteczniejszym preparatem jest Lek B, gdyż pacjenci w tej grupie statystycznie szybciej osiągają ustąpienie objawów choroby niż osoby leczone Lekiem A. Metoda Kaplana-Meiera oraz Fleminga-Harringtona podają zgodne (niższe wartości dla leku B) oraz bardzo zbliżone (rozbieżność między wynikami nie jest większa niż 0.05), co potwierdza stabilność i wiarygodność uzyskanych wyników niezależnie od zastosowanej metody estymacji.

Znaczącą różnicę można dostrzec przy średniej liczonej z i bez ogona. Średnia bez ogona wydaje się być zaniżona, ponieważ badanie kończy się po roku, a w tym momencie wielu pacjentów wciąż chorowało. Zastosowanie korekty “z ogonem” pozwala oszacować, jak długo ci pacjenci by jeszcze chorowali, dlatego wynik w tej wersji jest prawie dwa razy wyższy, ale znacznie bliższy prawdy o rzeczywistej skuteczności leczenia.

## 3 Lista 7

### 3.1 Zadanie 1

Celem zadania jest napisanie deklaracji funkcji o argumentach: dane (cenzurowane losowo), wartość poziomu ufności oraz wartość  $\tau$ . Funkcja ma za zadanie wyliczyć dolną oraz górną granicę przedziału ufności dla wartości średniej czasu życia.

#### 3.1.1 Wstęp teoretyczny

##### *Estymacja przedziałowa średniej czasu przeżycia*

**3.1.1.1 Estymator średniej** Estymator średniej  $\mu_\tau$  definiujemy jako pole pod krzywą na estymatora funkcji przeżycia  $\hat{S}(t)$  na przedziale od 0 do  $\tau$  wyraża się wzorem:

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t) dt \quad (4)$$

Wartość  $\tau$  powinna być dobrana tak, aby odzwierciedlała wiedzę a priori o maksymalnym czasie wystąpienia zdarzenia.

**3.1.1.2 Estymator wariancji** Oszacowanie estymatora wariancji wyraża się wzorem:

$$\hat{V}(\hat{\mu}_\tau) = \sum_{i=1}^D \left[ \int_{s_i}^\tau \hat{S}(t) dt \right]^2 \frac{d_i}{r_i(r_i - d_i)}. \quad (5)$$

gdzie:

- $D$  – liczba unikalnych czasów, w których wystąpiły zdarzenia niecenzurowane,
- $s_i$  – uporządkowane czasy wystąpienia zdarzeń niecenzurowanych,
- $r_i$  – liczebność zbioru narażonych (risk set) w czasie  $s_i$ ,
- $d_i$  – liczba zdarzeń w czasie  $s_i$ ,

\*sumowanie odbywa się jedynie po czasach  $s_i$ , w których zaobserwowano zdarzenia kompletne (niecenzurowane).

**3.1.1.3 Przedziały ufności** Dolna i górna granica przedziału ufności dla średniego czasu życia przy ucięciu w punkcie  $\tau$  wyrażają się wzorami:

$$T_L = \hat{\mu}_\tau - z(1 - \alpha/2)\sqrt{\hat{V}(\hat{\mu}_\tau)}, \quad (6)$$

$$T_U = \hat{\mu}_\tau + z(1 - \alpha/2)\sqrt{\hat{V}(\hat{\mu}_\tau)}. \quad (7)$$

gdzie:

- $z(q)$  – kwantyl rzędu  $q$  standardowego rozkładu normalnego,

### 3.1.2 Kod

Idea poniższej implementacji funkcji do wyznaczania przedziałów ufności zakłada wyznaczenie średniego czasu przeżycia ograniczonego przez  $\tau$  poprzez obliczenie pola powierzchni pod krzywą Kaplana-Meiera metodą sumowania prostokątów (analogicznie do podejścia z Listy 6). Wariancję estymatora obliczamy na podstawie znanego wzoru, iterując po czasach zdarzeń, gdzie wymagane całki wyznaczamy poprzez różnicę pól. Ostateczne granice przedziału ufności wyznaczamy przy założeniu asymptotycznej normalności rozkładu estymatora.

```
#implementacja przedziału ufności
przedzial_ufnosci <- function(dane, poziom_ufnosci, tau){
  KM_fit <- survfit(Surv(czas, status)~1, data = dane, type="kaplan-meier")

  czasy <- KM_fit$time
  prawd <- KM_fit$surv

  #musimy policzyć pole pod wykresem od 0 do tau
  pole_do_czasu <- function(czas_graniczny) {
    indeksy <- which(czasy <= czas_graniczny)
    #doklejamy 0 i 1
    os_czasu <- c(0, czasy[indeksy])
    os_y <- c(1, prawd[indeksy])

    if (tail(os_czasu, 1) < czas_graniczny) {
      os_czasu <- c(os_czasu, czas_graniczny)
      os_y <- c(os_y, tail(os_y, 1))}

    #pętla do liczenia pól prostokątów jak na liście 6
    pole <- 0
    liczba_punktow <- length(os_czasu)

    for (i in 1:(liczba_punktow - 1)) {
      szerokosc <- os_czasu[i+1] - os_czasu[i]
      wysokosc <- os_y[i]
      pole <- pole + (szerokosc * wysokosc)
    }
    return(pole)
  }

  #szukamy momentów, w których zaszło zdarzenie przed czasem tau
  #n.event > 0 oznacza zdarzenie, a nie cenzurę
  zdarzenie <- (KM_fit$n.event > 0) & (czasy <= tau)
  czasy_zdarzen <- czasy[zdarzenie] #s_i
```

```

liczba_zgonow <- KM_fit$n.event[zdarzenie]      #d_i
liczba_zagrozonych <- KM_fit$n.risk[zdarzenie]  #r_i

suma_wariancji <- 0

for (i in seq_along(czasy_zdarzen)) {
  s_i <- czasy_zdarzen[i]

  #obliczanie całki od s_i do tau.
  #Całka(s -> tau) = Całka(0 -> tau) - Całka(0 -> s)
  calka_od_s_do_tau <- pole_do_czasu(tau) - pole_do_czasu(s_i)

  #jak we wzorze na wariancje
  mianownik <- liczba_zagrozonych[i]*(liczba_zagrozonych[i] -
                                       liczba_zgonow[i])

  if (mianownik > 0) {
    skladnik <- (calka_od_s_do_tau^2) * (liczba_zgonow[i]/mianownik)
    suma_wariancji <- suma_wariancji + skladnik
  }
}
wariancja <- suma_wariancji
pierwiastek_z_war <- sqrt(suma_wariancji)

#przedziały ufności
alfa <- 1 - poziom_ufnosci
z <- qnorm(1 - alfa / 2)
dolna_granica <- pole_do_czasu(tau) - z * pierwiastek_z_war
gorna_granica <- pole_do_czasu(tau) + z * pierwiastek_z_war

return(c(L = dolna_granica, U = gorna_granica))
}

```

## 3.2 Zadanie 2

Celem zadania jest wyznaczenie realizacji przedziałów ufności dla średniego czasu do progresji choroby w dwóch grupach pacjentek (niski, wysoki), używając funkcji z zadania 3.1 dla dwóch wybranych wartości  $\tau$ .

### 3.2.1 Kod

```

#niski stopień zaawansowania
czas1 <- c(28, 89, 175, 195, 309, 377, 393, 421, 447,
          462, 709, 744, 770, 1106, 1206)
delta1 <- c(1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0)

```

```

#wysoki stopień zaawansowania
czas2 <- c(34, 88, 137, 199, 280, 291, 299, 300, 309, 351, 358,
          369, 369, 370, 375, 382, 392, 429, 451, 1119)
delta2 <- c(1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0)

dane <- data.frame(
  czas = c(czas1, czas2),
  status = c(delta1, delta2),
  progresja = factor(c(rep("Niski", length(czas1)),
                      rep("Wysoki", length(czas2))))
)

dane_niski <- subset(dane, progresja == "Niski")
dane_wysoki <- subset(dane, progresja == "Wysoki")

#tau = 1500
tau_niski_1500 <- przedzial_ufnosci(dane_niski,
                                    poziom_ufnosci = 0.95, tau = 1500)
tau_wysoki_1500 <- przedzial_ufnosci(dane_wysoki,
                                     poziom_ufnosci = 0.95, tau = 1500)

#tau = 1100
tau_niski_1100 <- przedzial_ufnosci(dane_niski,
                                    poziom_ufnosci = 0.95, tau = 1100)
tau_wysoki_1100 <- przedzial_ufnosci(dane_wysoki,
                                     poziom_ufnosci = 0.95, tau = 1100)

```

Tabela 3: Przedziały ufności

Tau	Stopień zaawansowania	Dolna granica	Górna granica
1500	Niski	589.9098	1285.5569
1500	Wysoki	232.7334	593.6749
1100	Niski	477.8077	953.2145
1100	Wysoki	254.6286	501.7797

### 3.2.2 Interpretacja

Analiza wyznaczonych przedziałów ufności (przedstawionych w tabeli 3) jednoznacznie wskazuje, że pacjentki z niskim stopniem zaawansowania choroby mają znacznie lepsze rokowania - ich średni czas bez progresji jest niemal dwukrotnie dłuższy niż w grupie wysokiego ryzyka. Przedziały ufności dla obu grup nieznacznie (bardzo delikatnie) na siebie nachodzą (są niemalże rozłączne), co wyraźnie potwierdza silną różnicę w przebiegu choroby pomiędzy grupami.

Porównując wyniki dla poszczególnych wartości  $\tau$  możemy dojść do wniosku, że mniejsza wartość  $\tau$  powoduje zwężenie (zmniejszenie się) przedziału ufności. Dodatkowo, możemy zasugerować, że w naszym przypadku bardziej odpowiednią wartością  $\tau$  jest 1100, gdyż mieści się ona w zakresie rzeczywistych obserwacji, podczas gdy estymacje dla  $\tau = 1500$  mogą nie odzwierciedlać prawidłowo rzeczywistości, gdyż wykraczają one poza czas, w którym dysponujemy danymi o pacjentkach.

## 4 Lista 8

### 4.1 Zadanie 1

Celem zadania było zweryfikowanie hipotezy zerowej

$$H_0 : h_1(t) = h_2(t) \quad \text{dla każdego } t,$$

czyli hipotezy o jednakowym rozkładzie czasu do progresji choroby w dwóch badanych grupach pacjentek, na poziomie istotności  $\alpha = 0.05$ , dla danych z zadania 3.2.

#### Testy

Do porównania rozkładów przeżycia wykorzystano testy oparte na estymatorze Nelsona-Aalena funkcji hazardu. W praktyce różne testy powstają poprzez zastosowanie innych funkcji wagowych  $W(t)$ :

- **Test log-rank:**

$$W(t_i) = 1, \quad i = 1, \dots, D$$

- **Test Gehana-Breslowa:**

$$W(t_i) = r_i, \quad i = 1, \dots, D$$

- **Test Tarone-Ware:**

$$W(t_i) = \sqrt{r_i}, \quad i = 1, \dots, D$$

- **Test Peto-Peto:**

$$W(t_i) = \tilde{S}(t_i) := \prod_{t_j \leq t_i} \left( 1 - \frac{d_j}{r_j + 1} \right)$$

gdzie:

- $d_{ij}$  - liczba zdarzeń w czasie  $t_i$  z próby  $X_j$ ,
- $r_{ij}$  - liczba jednostek w stanie ryzyka w czasie  $t_i$  w  $j$ -tej grupie,
- $d_i = \sum_{j=1}^k d_{ij}$  - łączna liczba zdarzeń w chwili  $t_i$ ,
- $r_i = \sum_{j=1}^k r_{ij}$  - łączna liczba jednostek w ryzyku w chwili  $t_i$ ,
- $D$  - liczba zdarzeń niecenzurowanych we wszystkich obserwowanych próbach.

#### Kod

Wykorzystano funkcję `logrank.test` z biblioteki `coin`, która umożliwia przeprowadzenie wszystkich wymienionych testów. Dla każdego testu obliczono odpowiadający im poziom istotności p-value.

```

test_log_rank <- logrank_test(Surv(czas, status) ~ progresja,
                             data = dane, type = "logrank")
test_gehan_breslowa <- logrank_test(Surv(czas, status) ~ progresja,
                                    data = dane, type = "Gehan-Breslow")
test_tarone_ware <- logrank_test(Surv(czas, status) ~ progresja,
                                 data = dane, type = "Tarone-Ware")
test_peto_peto <- logrank_test(Surv(czas, status) ~ progresja,
                               data = dane, type = "Peto-Peto")

wyniki <- data.frame(
  Test = c("Log-rank", "Gehan-Breslow", "Tarone-Ware", "Peto-Peto"),
  "p-value" = c(
    round(pvalue(test_log_rank), 4),
    round(pvalue(test_gehan_breslowa), 4),
    round(pvalue(test_tarone_ware), 4),
    round(pvalue(test_peto_peto), 4)),
  check.names = FALSE
)

```

Tabela 4: Wartości poziomów krytycznych

Test	p-value
Log-rank	0.0194
Gehan-Breslow	0.1336
Tarone-Ware	0.0552
Peto-Peto	0.0978

## Interpretacja

Wyniki zestawiono w tabeli 4. Decyzja o odrzuceniu  $H_0$  zapada, gdy wartość p-value jest mniejsza niż  $\alpha = 0.05$ . Zauważalne są różnice pomiędzy wartościami p-value uzyskanymi w poszczególnych testach.

Test log-rank, który przypisuje jednakowe wagi wszystkim zdarzeniom, zwraca p-value 0.0194, co pozwala na odrzucenie hipotezy zerowej o jednakowych rozkładach czasu do progresji choroby w obu grupach. Oznacza to, że log-rank wykrywa różnice pomiędzy grupami. Pozostałe testy Gehana-Breslowa, Tarone'a-Ware'a oraz Peto-Peto dają większe wartości p-value. Dla wszystkich tych testów wartość jest większa niż  $\alpha = 0.05$ , w związku z czym nie ma podstaw do odrzucenia  $H_0$ . Według tych testów, nie ma wystarczających dowodów, aby stwierdzić różnicę w rozkładach czasu do progresji choroby między grupami. Widać jednak, że różnica między p-value testu Tarone'a-Ware'a a log-rank jest nie aż tak wielka, większa zaś w testach Gehana-Breslowa i Peto-Peto.

## 4.2 Zadanie 2

W ramach zadania 2 naszkicowano wykres estymatorów Kaplana-Meiera funkcji przeżycia dla czasu do progresji choroby w dwóch badanych grupach pacjentek.

Dodatkowo, przedstawiono wykres unormowanych funkcji wagowych  $W(t_i)$  stosowanych w statystykach testowych rozważanych w zadaniu 4.1. Poniżej zamieszczono również fragment kodu ilustrujący sposób, w jaki obliczano te wagi.

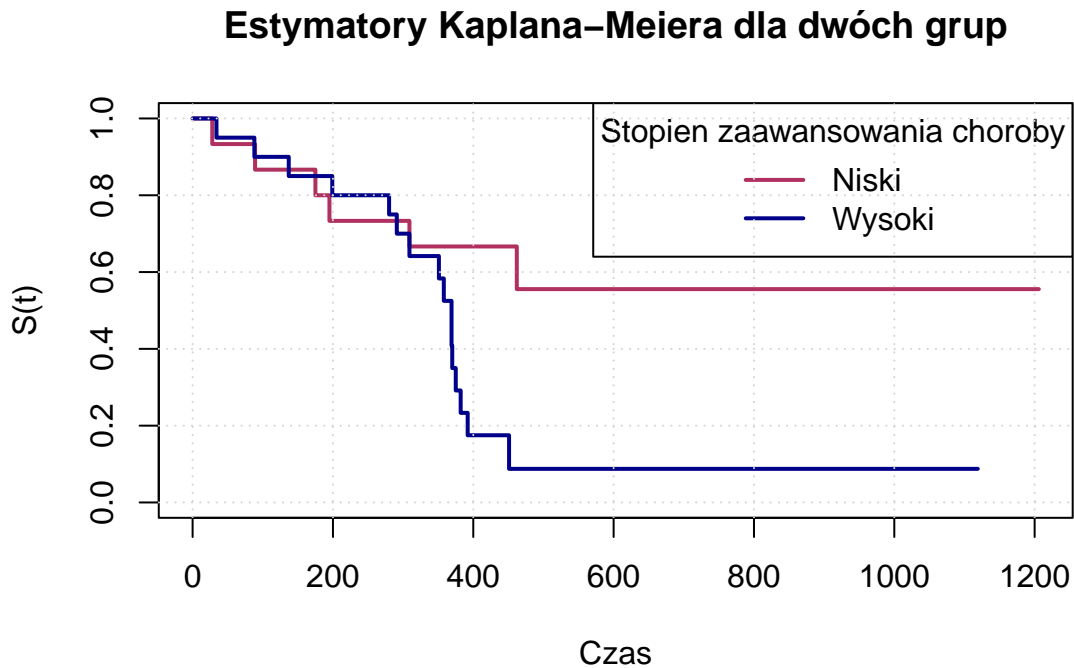
```
km_fit <- survfit(Surv(czas, status) ~ 1, data = dane)

czasy <- sort(unique(dane$czas[dane$status == 1]))
D <- length(czasy)
r_i <- numeric(D)
d_i <- numeric(D)

for (i in 1:D) {
  czas <- czasy[i]
  r_i[i] <- sum(dane$czas >= czas)
  d_i[i] <- sum(dane$czas == czas & dane$status == 1)
}

wagi_log_rank <- rep(1, D)
wagi_gehan_breslow <- r_i
wagi_tarone_ware <- sqrt(r_i)
wagi_peto_peto <- cumprod(1 - (d_i/(r_i+1)))

wagi_log_rank_norm <- wagi_log_rank / sum(wagi_log_rank)
wagi_gehan_breslow_norm <- wagi_gehan_breslow / sum(wagi_gehan_breslow)
wagi_tarone_ware_norm <- wagi_tarone_ware / sum(wagi_tarone_ware)
wagi_peto_peto_norm <- wagi_peto_peto / sum(wagi_peto_peto)
```



Rysunek 7: Estymatory Kaplana-Meiera dla dwóch grup pacjentów o różnym stopniu zaawansowania choroby.

## Interpretacja

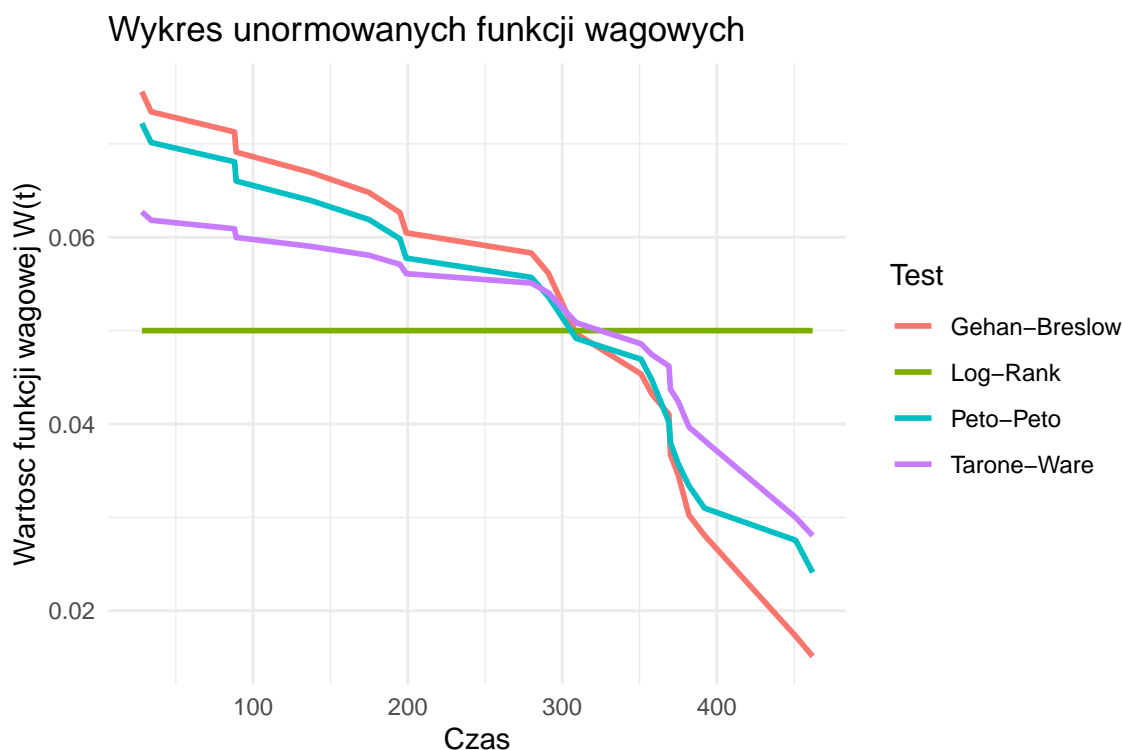
Rysunek 7 przedstawia estymatory Kaplana-Meiera funkcji przeżycia  $S(t)$  dla dwóch grup pacjentek różniących się stopniem zaawansowania choroby. Oś pozioma przedstawia czas do progresji choroby, natomiast oś pionowa szacowane prawdopodobieństwo przeżycia bez progresji do danego momentu.

Krzywe zostały wyznaczone na podstawie funkcji `survfit(Surv(czas, status) ~ progresja, data = dane)` i pokazują przebieg funkcji przeżycia w obu grupach.

Krzywa dla grupy o wysokim stopniu zaawansowania choroby spada szybciej, co oznacza, że pacjentki mają niższe prawdopodobieństwo przeżycia bez progresji w późniejszym okresie. Widoczny jest gwałtowny spadek w okolicy dnia 400, co wskazuje na liczne nawroty choroby zachodzące w krótkich odstępach czasu.

Natomiast krzywa dla grupy o niskim stopniu zaawansowania choroby pozostaje wyżej przez większą część okresu. W grupie tej progresja zachodzi rzadziej i później, obserwujemy dłuższe okresy stabilnego przebiegu. Funkcja przeżycia utrzymuje się na wyższych wartościach.

Krzywe zaczynają się rozchodzić już na dosyć wczesnym etapie obserwacji, co pokazuje, że różnice między grupami pojawiają się od początku badania.



Rysunek 8: Unormowane funkcje wagowe dla testów Log-Rank, Gehan-Breslow, Tarone-Ware i Peto-Peto.

## Interpretacja

Na rysunku 8 przedstawiono przebieg unormowanych funkcji wagowych  $W(t_i)$  stosowanych w czterech testach: Log-Rank, Gehan-Breslow, Tarone-Ware oraz Peto-Peto.

Linia odpowiadająca testowi log-rank jest stała w czasie, co oznacza, że wszystkie zdarzenia są traktowane jednakowo, niezależnie od momentu ich wystąpienia. Wagi w tym teście nie faworyzują ani wczesnych, ani późnych zdarzeń.

Funkcja wagowa dla testu Gehana-Breslowa przyjmuje największe wartości na początku i systematycznie maleje wraz z upływem czasu. Oznacza to, że ten test nadaje największe znaczenie zdarzeniom występującym na początku.

Krzywa odpowiadająca testowi Peto-Peto również ma charakter malejący, jej przebieg jest trochę bardziej wygładzony niż test Gehana-Breslowa.

Test Tarone'a-Ware'a ma przebieg pośredni pomiędzy testem log-rank a testem Gehana-Breslowa. Jego wagi maleją łagodniej w czasie, przez co bardziej uwzględnia zarówno wczesne, jak i późniejsze różnice pomiędzy grupami.

Wykres pokazuje więc, że poszczególne testy różnią się sposobem ważenia zdarzeń w czasie, co wpływa na ich czułość na różnice pomiędzy grupami w różnych fazach obserwacji.

## Podsumowanie

Różnice w wartościach p-value przedstawione w zadaniu 4.1 w tabeli 4 wynikają bezpośrednio z tego, które momenty obserwacji są akcentowane przez funkcje wagowe. Log-rank, traktując wszystkie zdarzenia jednakowo, wykrył istotną różnicę między grupami. Testy akcentujące początek obserwacji jak Gehan-Breslow i Peto-Peto nie wykazały istotności, ponieważ w tym okresie różnice między krzywymi przeżycia były mniej wyraźne. Tarone-Ware, jako kompromis, dał wynik pośredni.

Na rysunku 7 widać że krzywe Kaplana-Meiera dla dwóch grup pacjentek zaczynają się rozchodzić dopiero w późniejszym okresie obserwacji co oznacza że różnice w przeżyciu ujawniają się głównie w dalszym okresie. Test log-rank nadaje równą wagę wszystkim zdarzeniom niezależnie od czasu dlatego był najbardziej czuły na różnice, które pojawiły się później i jako jedyny wykazał istotność.