Analiza Przeżycia

Raport 2

Wiktor Niedźwiedzki (258882)
Filip Michewicz (282239)

8 grudnia 2025 Anno Domini

Spis treści

Spis wykresów

Spis tabel

1 Lista 5

Lista obejmuje estymatory funkcji przeżycia dla danych niecenzurowanych, w szczególności estymator Kaplana–Meiera oraz estymator Fleminga–Harringtona, zarówno bez korekty, jak i z ogonem estymowanym według propozycji Browna, Hollandera i Kowara. Dodatkowo przeprowadzono oszacowanie wartości funkcji przeżycia w chwili cenzurowania oraz w dwukrotności czasu cenzurowania, na podstawie symulacji dla danych generowanych z uogólnionego rozkładu wykładniczego $\mathcal{GE}(\lambda, \alpha)$.

1.1 Zadanie 1

W tym zadaniu wygenerowano wykres estymatorów Kaplana–Meiera oraz Fleminga–Harringtona funkcji przeżycia dla danych z zadania 3 z listy 2, których opis przedstawiono w poprzednim raporcie.

1.1.1 Estymator Kaplana-Meiera

Estymator Kaplana–Meiera jest nieparametrycznym estymatorem funkcji przeżycia, opartym na iloczynie warunkowych prawdopodobieństw przeżycia kolejnych chwil zdarzeń. Definiuje się go jako

$$\hat{S}(t) = \prod_{i:\, t_{(i)} \le t} \left(1 - \frac{d_i}{r_i}\right),$$

gdzie

$$r_i = \sum_{j=1}^{n} \mathbf{1}_{[t_{(i)}, \infty)}(t_j), \qquad d_i = \sum_{j=1}^{n} \mathbf{1}_{\{t_{(i)}\}}(t_j)\, \mathbf{1}_{\{1\}}(\delta_j).$$

Wielkość r_i to liczba obserwacji, które w chwili $t_{(i)}$ pozostają w stanie ryzyka, czyli dotrwały do czasu $t_{(i)}$ bez wcześniejszego zdarzenia ani cenzurowania i mogą jeszcze doświadczyć zdarzenia w tym momencie, natomiast d_i to liczba zdarzeń (niezależnych od cenzurowania) zachodzących dokładnie w czasie $t_{(i)}$.

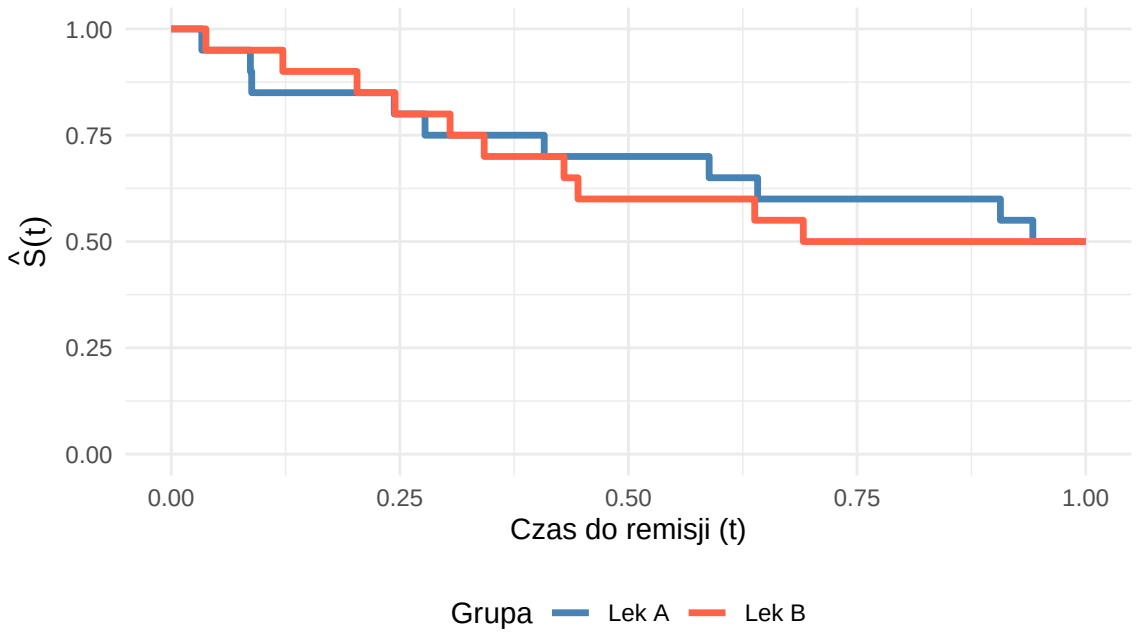
Estymator Kaplana–Meiera nie został zdefiniowany ręcznie w tym raporcie, lecz jego implementacja znajduje się w bibliotece `survival` w pakiecie R. Funkcja `survfit` domyślnie dopasowuje dane w formacie `Surv`, czyli takim, w którym określone są czasy obserwacji oraz wskaźniki cenzurowania, i domyślnie używa estymatora Kaplana–Meiera.

Poniżej przedstawiono blok kodu w R, który realizuje estymację funkcji przeżycia za pomocą tego estymatora:

```
surv.A <- Surv(df.A$times, df.A$deltas)
surv.B <- Surv(df.B$times, df.B$deltas)

# Estymator Kaplana-Meiera
fit.KM.A <- survfit(surv.A ~ 1)
fit.KM.B <- survfit(surv.B ~ 1)

# Ramki danych do wykresu
plot.A <- data.frame(time = c(0, fit.KM.A$time),
                     surv = c(1, fit.KM.A$surv),
                     group = "A")

plot.B <- data.frame(time = c(0, fit.KM.B$time),
                     surv = c(1, fit.KM.B$surv),
                     group = "B")
```

```r
plot.df  <- rbind(plot.A, plot.B)
```

Poniżej przedstawiono wykres estymowanej funkcji przeżycia uzyskanej przy użyciu estymatora Kaplana–Meiera dla danych z zadania 3 z listy 2.



Wykres 1: Estymator Kaplana-Meiera dla danych dotyczących leków

Wykres 1. przedstawia estymowaną funkcję przeżycia za pomocą estymatora Kaplana–Meiera. Dane w obu przypadkach są prawostronnie cenzurowane typu I, zatem największe wartości w zbiorach danych są cenzurowane, a po ostatnim pełnym (niecenzurowanym) czasie występują obserwacje cenzurowane, przez co estymator nie jest określony dla $t > t_{(n)}$. Dlatego do czasu $t \leq t_{(n)}$ estymator Kaplana–Meiera jest dobrze określony. W przypadku gdyby obserwacja $t_{(n)}$ była kompletna, to dla $t > t_{(n)}$ mielibyśmy $\hat{S}(t) = 0$.

1.1.2 Estymator Fleminga-Harringtona

Estymator Fleminga–Harringtona jest estymatorem funkcji przeżycia, opartym na zależności między funkcją przeżycia a funkcją skumulowanego hazardu; wykorzystuje estymator funkcji hazardu, a w konsekwencji skumulowanej funkcji hazardu zaproponowany przez Nelsona–Aalena. Szczegóły dotyczące konstrukcji estymatora i jego relacji do funkcji skumulowanego hazardu przedstawiono na wykładzie. Funkcję przeżycia estymuje się według wzoru:

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

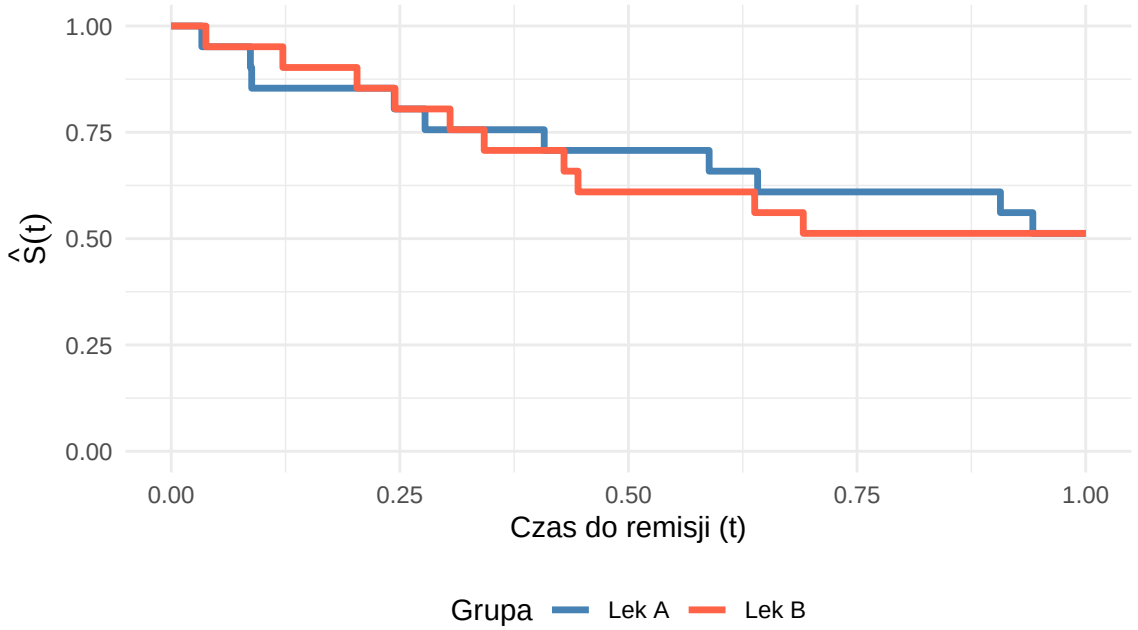
gdzie d_j to liczba zdarzeń w chwili $t_{(j)}$, a r_j - liczba obserwacji pozostających w stanie ryzyka w czasie $t_{(j)}$.

Podobnie jak w poprzednim podpunkcie, estymator Fleminga–Harringtona nie został zdefiniowany ręcznie w tym raporcie, lecz jego implementacja znajduje się w bibliotece `survival` w pakiecie R. Tym razem w funkcji `survfit` zastosowano opcję `type = "fleming-harrington"`.

Poniżej przedstawiono blok kodu w R, który realizuje estymację funkcji przeżycia za pomocą tego estymatora:

```r
# Estymator Fleminga-Harringtona
fit.FH.A <- survfit(surv.A ~ 1, type = "fleming-harrington")
fit.FH.B <- survfit(surv.B ~ 1, type = "fleming-harrington")

# Ramki danych do wykresu
plot.A <- data.frame(time = c(0, fit.FH.A$time),
                     surv = c(1, fit.FH.A$surv),
                     group = "A")

plot.B <- data.frame(time = c(0, fit.FH.B$time),
                     surv = c(1, fit.FH.B$surv),
                     group = "B")

plot.df  <- rbind(plot.A, plot.B)
```

Poniżej przedstawiono wykres estymowanej funkcji przeżycia uzyskanej przy użyciu estymatora Fleminga-Harringtona.



Wykres 2: Estymator Fleminga-Harringtona dla danych dotyczących leków

Wykres 2. przedstawia estymowaną funkcję przeżycia za pomocą estymatora Fleminga-Harringtona. Uwagi dotyczące bycia dobrze określonym są analogiczne jak do estymatora Kaplana-Meiera.

Obydwa estymatory dają podobne wyniki. W krótkim czasie lek B wydaje się korzystniejszy (wyższa funkcja przeżycia, dłuższy czas do remisji), natomiast dla $t > 0,5$ bardziej efektywny staje się lek A. Do dokładniejszej analizy warto zastosować, np. test Kołmogorowa–Smirnowa.

1.2 Zadanie 2

Zadanie dotyczy generowania wykresu funkcji przeżycia estymowanej metodą Kaplana–Meiera, z uwzględnieniem oszacowania ogona funkcji przeżycia, zaproponowanego przez Browna, Hollandera i Kowara.

Brown, Hollander i Kowar (1974) sugerowali oszacowanie ogona funkcji przeżycia przez odpowiednio dobraną funkcję przeżycia rozkładu wykładniczego, taką, aby w punkcie t^+ (ostatnim zarejestrowanym czasie) była równa $S(t^+)$.

Po krótkich obliczeniach przeprowadzonych na wykładzie otrzymuje się:

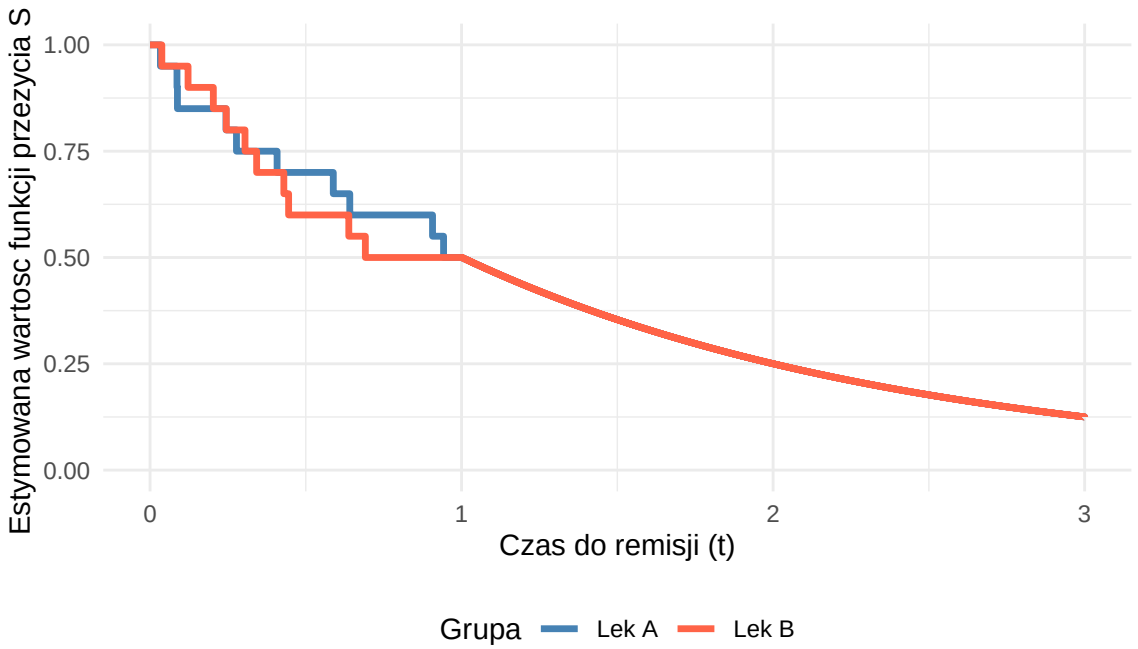
$$\hat{S}(t) = \exp\left(\frac{\ln \hat{S}(t^+)}{t^+} t\right), \quad t > t^+.$$

Dla $t \leq t^+$ stosuje się klasyczny estymator Kaplana–Meiera.

Poniżej przedstawiono blok kodu w R, który umożliwia estymację funkcji przeżycia wraz z uwzględnieniem ogona.

```r
BHK.tail <- function(df, end = 3, type = "kaplan-meier") {
  # Estymator KM
  fit <- survfit(Surv(df$times, df$deltas) ~ 1, type = type)
  df.complete <- data.frame(time = fit$time,
                            surv = fit$surv)

  # Ogin
  t_plus <- tail(df.complete$time, 1)
  S_plus <- tail(df.complete$surv, 1)

  theta <- -log(S_plus) / t_plus

  # Generowanie ogona
  time_tail <- seq(t_plus, end, length.out = 1000)
  surv_tail <- exp(-theta * time_tail)
  df_tail <- data.frame(time = time_tail, surv = surv_tail)

  df.complete <- rbind(df.complete, df_tail)

  # Dodanie punktu początkowego
  df.complete <- rbind(data.frame(time = 0, surv = 1), df.complete)
  return(df.complete)
}
```

Poniżej przedstawiono wykres estymowanej funkcji przeżycia uzyskanej przy użyciu estymatora Kaplana–Meiera wraz z ogonem dla danych z zadania 3 z listy 2.



Wykres 3: Estymator Kaplana-Meiera wraz z ogonem Browna, Hollandera i Kowara dla danych dotyczących leków

Wykres 3. przedstawia estymator Kaplana–Meiera wraz z ogonem zaproponowanym przez Browna, Hollandera i Kowara dla danych dotyczących leków. Do czasu ostatniej obserwacji kompletnej wykorzystano klasyczny estymator Kaplana–Meiera, natomiast dla czasów przekraczających tę obserwację wartość funkcji przeżycia (tzw. ogon) estymowana jest przy użyciu odpowiednio dobranej funkcji wykładniczej. Warto zauważyć, że estymowana funkcja jest ciągła w punkcie t^+. Ze względu na to, że $t_A^+ = t_B^+ = t_0 = 1$, ogony w obu grupach pokrywają się.

1.3 Zadanie 3

Zadanie polega na wygenerowaniu $M = 1000$ zbiorów danych cenzurowanych I-go typu z uogólnionego rozkładu wykładniczego $\mathcal{GE}(\lambda, \alpha)$ dla $\alpha = 2$ i $\lambda = 2$, przy licznościach próby $n = 30, 50, 100$. Wartość t_0 przyjęto równą wartości oczekiwanej rozkładu $\mathcal{GE}(\lambda, \alpha)$ [1]:

$$t_0 = \lambda \left(\psi(\alpha + 1) - \psi(1)\right)$$

gdzie $\psi(\cdot)$ jest funkcją digamma, tj. logarytmiczną pochodną funkcji gamma.

Na podstawie wygenerowanych zbiorów danych oszacowano wartość funkcji przeżycia w punktach t_0 i $2t_0$, korzystając z estymatora Kaplana–Meiera z uwzględnieniem ogona estymowanego zgodnie z propozycją Browna, Hollandera i Kowara. Ostatecznie wygenerowano histogramy oszacowań w tych punktach dla każdego n i oceniono, czy istnieją podstawy do przypuszczenia, że estymator Kaplana–Meiera jest asymptotycznie normalny.

Konstrukcja estymatora Kaplana–Meiera wraz z "ogonem" została omówiona w dwóch poprzednich zadaniach i nie będzie tutaj ponownie przytaczana.

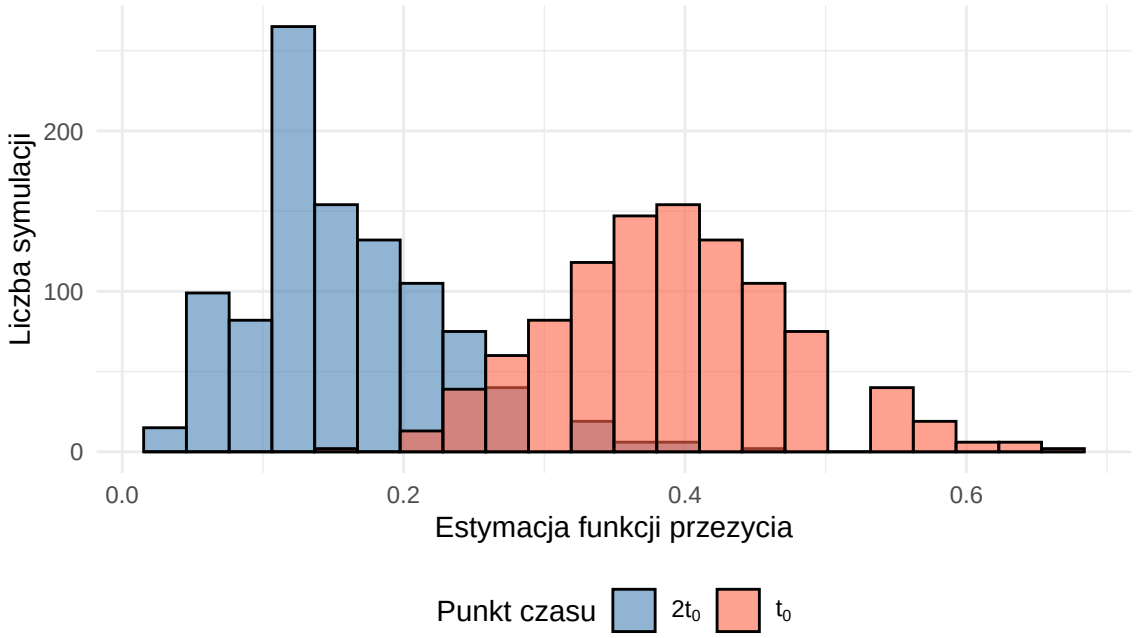
Poniżej przedstawiono blok kodu w pakiecie R wyznaczający, za pomocą estymatora Kaplana–Meiera z "ogonem", estymowaną wartość funkcji przeżycia w punkcie t.

```r
KM.surv.value <- function(df, type = "kaplan-meier", t) {
  # Estymator KM
  fit <- survfit(Surv(df$times, df$deltas) ~ 1, type = type)
  event_idx <- which(fit$n.event > 0)
  df.complete <- data.frame(time = fit$time[event_idx],
                            surv = fit$surv[event_idx])

  df.complete <- data.frame(time = fit$time, surv = fit$surv)

  t_plus <- tail(df.complete$time, 1)

  if (t <= t_plus) {
    value <- df.complete$surv[max(which(df.complete$time <= t))]
  } else {
    S_plus <- tail(df.complete$surv, 1)

    theta <- -log(S_plus) / t_plus

    value <- exp(-theta * t)
  }
  return(value)
}
```

Poniżej przedstawiono blok kodu w pakiecie R przeprowadzający opisaną na początku zadania symulację. W trakcie obliczeń zapisywane są estymowane wartości funkcji przeżycia w punktach t_0 oraz $2t_0$. Na podstawie wyników sporządzono stosowne histogramy.

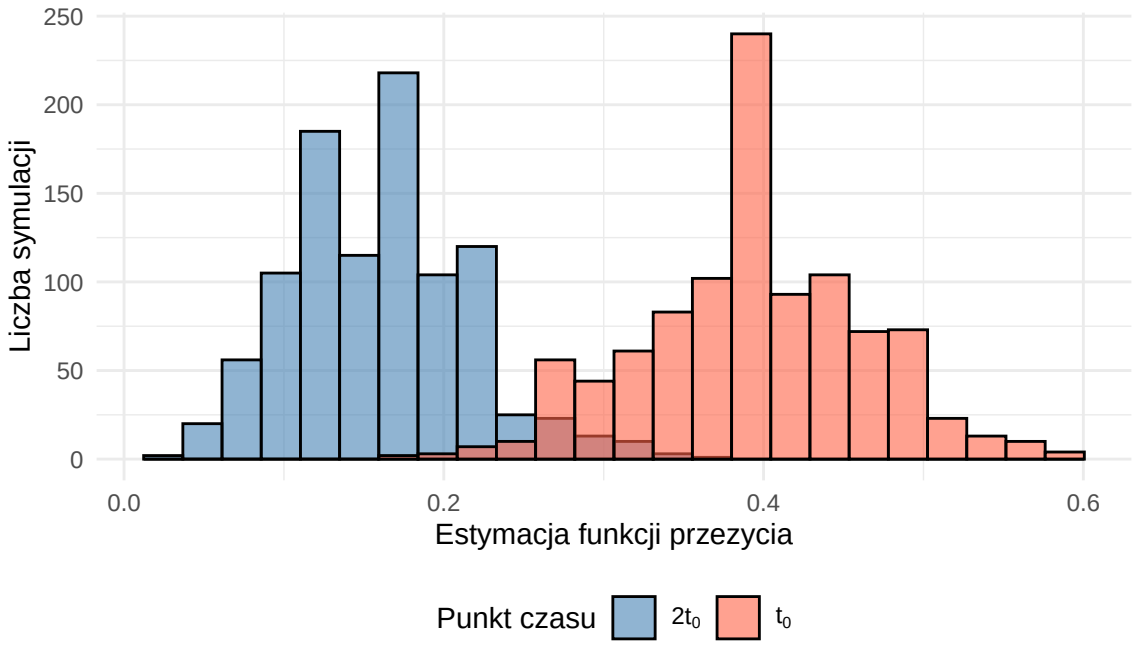
```r
M <- 1000
n_values <- c(30, 50, 100)
alpha <- 2
lambda <- 2

t0 <- (digamma(alpha + 1) - digamma(1)) / lambda

results_t0 <- list()
results_2t0 <- list()

for (n in n_values) {
  surv_t0 <- numeric(M)
  surv_2t0 <- numeric(M)

  m <- 0
  while (m < M) {
    # generowanie danych cenzurowanych I typu
    cenzurowane <- GE_cenzurowanie_I_typu(t0, alpha, lambda, n)
    if (sum(cenzurowane$deltas) == 0) {
      next
      }
    m <- m + 1

    # estymator KM z "ogonem" w t0
    surv_t0[m] <- KM.surv.value(cenzurowane, t = t0)
```

```r
    # estymator KM z "ogonem" w 2*t0
    surv_2t0[m] <- KM.surv.value(cenzurowane, t = 2*t0)
  }

  results_t0[[as.character(n)]] <- surv_t0
  results_2t0[[as.character(n)]] <- surv_2t0
}
```



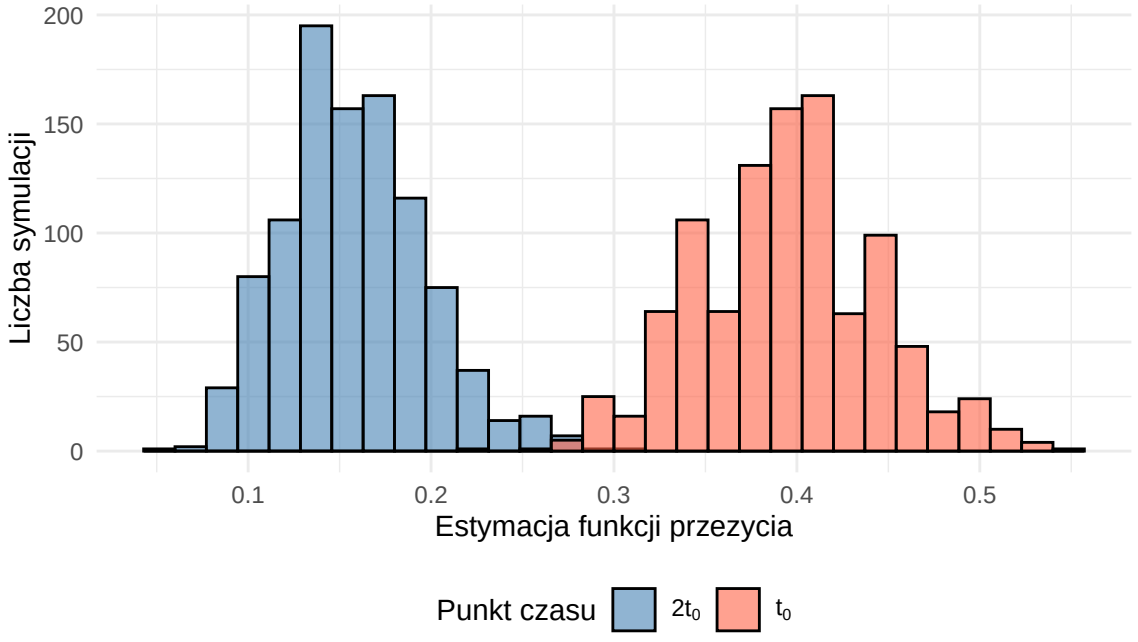
Wykres 4: Histogram oszacowań funkcji przeżycia w punktach t_0 oraz $2t_0$ - liczność próby n = 30

Wykres 4. przedstawia estymowaną wartość funkcji przeżycia w punktach t_0 oraz $2t_0$ dla próby o liczności $n = 30$. Estymowane wartości dla $2t_0$ są mniejsze niż dla t_0 co zgadza się z oczekiwaniami.



Wykres 5: Histogram oszacowań funkcji przeżycia w punktach t_0 oraz $2t_0$ - liczność próby n = 50

Wykres 5. przedstawia estymowaną wartość funkcji przeżycia w punktach t_0 oraz $2t_0$ dla próby o liczności $n = 50$. W porównaniu z próbą o liczności $n = 30$ przedstawionej na Wykresie 4. estymowane wartości przesunęły się ku mniejszym wartością.



Wykres 6: Histogram oszacowań funkcji przeżycia w punktach t_0 oraz $2t_0$ - liczność próby n = 100

Wykres 6. przedstawia estymowaną wartość funkcji przeżycia w punktach t_0 oraz $2t_0$ dla próby o liczności

$n = 30$. Podobnie jak na poprzednim Wykresie wraz ze wzrostem próby wartości przesunęły się ku mniejszym wartością.

Wraz ze wzrostem liczności próby, histogramy z poprzednich wykresów układają się w kształt dzwonowy charakterystyczny dla rozkładu normalnego. To z kolei pozwala przypuszczać, że estymator Kaplana–Meiera jest asymptotycznie normalny w punktach t_0 oraz $2t_0$. Do dalszej analizy wykorzystano test Shapiro-Wilka, który w pakiecie R dostępny jest w funckji `shapiro.test` z pakietu `stats`.

Tabela 1: Wartości p testu Shapiro-Wilka dla oszacowań funkcji przeżycia w punktach t_0 oraz $2t_0$

n	t_0	$2t_0$
30	0.000000015	0.0e+00
50	0.000036425	0.0e+00
100	0.003136258	1.2e-08

Z Tabeli 1. można odczytać że dla wszystkich przypadku p-value jest równe zeru lub bliskie zera, a zatem dla prawie każdego poziomu α powinniśmy odrzucić hipotezę o normalności rozkładów, zarówno dla punktu t_0, jak i $2t_0$.

2 Lista 6

Lista dotyczy estymacji średniego czasu życia w oparciu o estymatory Kaplana–Meiera oraz Fleminga–Harringtona, z uwzględnieniem szacowania "ogona" zaproponowanego przez Browna, Hollandera i Kowara. Dodatkowo dokonano estymacji średniego czasu życia dla danych pochodzących z zadania 3 z listy 2.

2.1 Zadanie 1

Zadanie polega na zdefiniowaniu funkcji obliczającej średni czas życia w oparciu o estymatory Kaplana–Meiera oraz Fleminga–Harringtona, z uwzględnieniem ogona wykładniczego według propozycji Browna, Hollandera i Kowara. Ze względu na podobieństwo wyglądu wynikowego estymatora dla obu metod, obliczenia zostaną przeprowadzone w uogólnionym przypadku.

Wartość oczekiwaną zmiennej losowej X, absolutnie ciągłej względem miary Lebesgue'a, można oszacować na dwa sposoby:

1. Obliczając
$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x) dx,$$
gdzie $f(x)$ jest funkcją gęstości rozkładu zmiennej losowej (X).

2. Jeżeli rozkład X ma nośnik $(0, \infty)$, to obliczając:
$$\mathbb{E}[X] = \int_0^\infty S(x) dx,$$
czyli poprzez całkowanie ogona funkcji przeżycia $S(x)$.

Jako że zarówno estymator Kaplana–Meiera, jak i Fleminga–Harringtona są estymatorami funkcji przeżycia, wartość oczekiwaną zmiennej losowej można oszacować za pomocą wzoru:

$$\hat{\mu} = \int_0^\infty \hat{S}(x) dx,$$

gdzie $\hat{S}(x)$ oznacza estymowaną funkcję przeżycia.

Ze względu na to że $\hat{S}(x)$ na przedziale $(0, t^+)$ jest funkcją schodkową, dlatego $\hat{\mu}$ będzie obliczana według wzoru:

$$\hat{\mu} \quad = \sum_{i:, t_{(i)} \leq t^+} \hat{S}\left(t_{(i)}\right)\left(t_{(i)} - t_{(i-1)}\right), \quad t_{(0)} := 0,$$

gdzie $\hat{S}(t_{(i)})$ to wartość estymatora funkcji przeżycia w chwili $t_{(i)}$, a sumowanie jest wykonywane tylko do czasu ostatniej obserwacji t^+. Powyżej t^+ obydwa estymatory są nieokreślone.

Jeżeli natomiast estymator będzie zawierał "ogon" według propozycji Browna, Hollandera i Kowara należy powiększyć wartość μ o wartość oczekiwaną na tym właśnie "ogonie".

$$\begin{aligned}\int_{t^+}^\infty e^{-\theta t}\, dt &= \lim_{b\to\infty} \int_{t^+}^b e^{-\theta t}\, dt = \lim_{b\to\infty} \left[-\frac{1}{\theta} e^{-\theta t}\right]_{t^+}^b = \\ &= \lim_{b\to\infty} \left(-\frac{1}{\theta} e^{-\theta b} + \frac{1}{\theta} e^{-\theta t^+}\right) = \frac{e^{-\theta t^+}}{\theta} = \frac{S(t^+)\cdot t^+}{-\ln S(t^+)}\end{aligned}$$

Ostatecznie estymowana wartość oczekiwana obliczana jest ze wzoru:

$$\hat{\mu} \quad = \sum_{i:,t_{(i)} \leq t^+} \hat{S}\left(t_{(i)-1}\right)\left(t_{(i)} - t_{(i-1)}\right) - \frac{S(t^+) \cdot t^+}{\ln S(t^+)}, \quad t_{(0)} := 0, \quad \hat{S}\left(t_{(0)}\right) = 1$$

Poniżej przedstawiono blok kodu w R obliczający średni czas życia. Parametr `type` pozwala wybrać, czy obliczenia mają być wykonane dla estymatora Kaplana–Meiera czy Fleminga–Harringtona, natomiast parametr `tail` określa, czy ma być uwzględniony ogon wykładniczy według propozycji Browna, Hollandera i Kowara.

```r
expected.life <- function(df, type = "kaplan-meier", tail = FALSE) {
  fit <- survfit(Surv(df$times, df$deltas) ~ 1, type = type)

  df.complete <- data.frame(time = c(0, fit$time),
                            surv = c(1, fit$surv))

  # dyskretna całka
  dt <- diff(df.complete$time)
  surv_val <- head(df.complete$surv, -1)
  integral_KM <- sum(surv_val * dt)

  # ogon wykładniczy
  if (tail) {
    t_plus <- tail(df.complete$time, 1)
    S_plus <- tail(df.complete$surv, 1)
    theta <- -log(S_plus) / t_plus
    tail_integral <- S_plus / theta
    return(integral_KM + tail_integral)
  }

  return(integral_KM)
}
```

2.2 Zadanie 2

Zadanie polega na wykorzystaniu napisanej w poprzednim zadaniu funkcji do oszacowania średniego czasu do emisji choroby pacjentów leczonym lekiem A i lekiem B na podstawie danych z zadania 3 z listy 2.

Wyniki przedstawiono w tabelach poniżej.

Tabela 2: Średni czas życia estymowany metodami Kaplana-Meiera oraz Fleminga-Harringtona

	Lek A	Lek B
Kaplan-Meier	1.432	1.394
Fleming-Harrington	1.484	1.447

Tabela 2. przedstawia średni czas życia estymowany metodami Kaplana–Meiera oraz Fleminga–Harringtona. W obu grupach estymatory wskazują, że średni czas życia w grupie przyjmującej lek A jest większy niż w grupie przyjmującej lek B. Estymator Fleminga-Harringtona zwraca nieco wyższe wartości średniego czasu życia niż estymator Kaplana–Meiera, co wynika z samej konstrukcji estymatora i sposobu ważenia obserwacji cenzurowanych.

3 Lista 7

Lista polega na napisaniu funkcji która na podstawie danych cenzurowanych losowo oblicza dolną i górną granicę przedziału ufności dla średniego czasu życia na zadanym poziomie ufności $1-\alpha$ oraz dla określonej wartości τ, która odzwierciedla naszą wiedzę a priori o maksymalnym czasie do wystąpienia zdarzenia. Następnie, wykorzystując rzeczywiste dane z badania Mayo Clinic dotyczące czasu do progresji choroby w dwóch grupach pacjentek, zastosowano tę funkcję do wyznaczenia przedziałów ufności dla wybranych wartości τ i porównano otrzymane wyniki między grupami.

3.1 Zadanie 1

Zadanie polega na zdefiniowaniu funkcji, która dla zadanych danych, poziomu istotności α oraz wartości τ (odzwierciedlającej naszą wiedzę a priori o maksymalnym czasie do wystąpienia zdarzenia) zwraca dolną i górną granicę przedziału ufności dla średniego czasu życia, na poziomie ufności $1-\alpha$.

Do wyznaczenia przedziałów ufności korzystamy z twierdzenia że, jeśli F i G są dystrybuantami odpowiednio czasu życia i czasu cenzurowania, ciągłymi na przedziale $[0,T]$, a ponadto $F(T)<1$, to zachodzi zbieżność

$$\frac{\hat{\mu}_\tau - \mu}{\sqrt{\widehat{V}(\hat{\mu}_\tau)}} \xrightarrow{d} N(0,1).$$

Wynika z tego, że przedział ufności postaci $[T_L, T_U]$, gdzie

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

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

jest asymptotycznym punktowym przedziałem ufności dla wartości średniej μ rozkładu czasu życia na poziomie ufności $1-\alpha$. Zmienna $z(q)$ oznacza kwantyl rzędu q standardowego rozkładu normalnego.

Estymator $\hat{\mu}_\tau$ średniego czasu życia ograniczonego do punktu τ (restricted mean survival time) jest definiowany jako:

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

gdzie $\hat{S}(t)$ oznacza estymator Kaplan–Meiera funkcji przeżycia.

Wariancja estymatora $\hat{\mu}_\tau$ jest przybliżana za pomocą wzoru:

$$\widehat{V}(\hat{\mu}_\tau) = \sum_{i=1}^{D} \left(\int_{s_i}^{\tau} \widehat{S}(t)\,\mathrm{d}t \right)^2 \frac{d_i}{r_i(r_i-d_i)}.$$

gdzie

$$r_i = \sum_{j=1}^{n} \mathbf{1}_{[t_{(i)},\infty)}(t_j) \qquad d_i = \sum_{j=1}^{n} \mathbf{1}_{\{t_{(i)}\}}(t_j)\,\mathbf{1}_{\{1\}}(\delta_j),$$

a D oznacza liczbę zaobserwowanych (niecenzurowanych) zdarzeń, natomiast $s_i = t_{(i)}$ jest i-tą statystyką pozycyjną czasów zdarzeń.

Wielkość r_i to liczba jednostek znajdujących się w stanie ryzyka w chwili $t_{(i)}$, czyli obserwacji, które dotrwały do tego czasu bez wcześniejszego zdarzenia lub cenzurowania. Z kolei d_i to liczba (niecenzurowanych) zdarzeń występujących dokładnie w chwili $t_{(i)}$.

Poniżej przedstawiono blok kodu w języku R zawierający funkcję, która oblicza wartości $\int_{s_i}^{\tau} \hat{S}(t), dt$ oraz estymator wariancji $\widehat{V}(\hat{\mu}_\tau)$, a następnie wyznacza granicę dolną T_L i górną T_U asymptotycznego przedziału ufności dla średniego czasu życia, dla zadanego zbioru danych, poziomu istotności α oraz wartości τ.

```r
integrate.S <- function(df, a, b) {

  fit <- survfit(Surv(df$times, df$deltas) ~ 1)
  event_idx <- which(fit$n.event > 0)

  df.complete <- data.frame(
    time = c(0, fit$time[event_idx]),
    surv = c(1, fit$surv[event_idx])
  )

  Time <- df.complete$time
  Surv <- df.complete$surv

  # a za ostatnim skokiem - całkowanie po stałej wartości
  if (a >= tail(Time, 1))
    return((b - a) * tail(Surv, 1))

  # Pierwszy punkt > a
  t_idx_candidates <- which(Time > a)
  t_idx <- min(t_idx_candidates)
  prev_idx <- max(1, t_idx - 1) # KM(a) = wartość z poprzedniego skoku

  # pierwszy fragment całki: od a do najbliższego punktu siatki
  integral_S <- Surv[prev_idx] * (Time[t_idx] - a)

  # b przed ostatnim skokiem
  if (b <= tail(Time, 1)) {
    # ostatni punkt < b
    T_idx_candidates <- which(Time < b)
    T_idx <- max(T_idx_candidates)

    # fragment końcowy: od ostatniego punktu < b do b
    integral_S <- integral_S + Surv[T_idx] * (b - Time[T_idx])

    dt <- diff(df.complete$time)
    surv_val <- head(df.complete$surv, -1)

    # środkowa część sumy prostokątów
    if (T_idx - 1 >= t_idx)
      integral_S <- integral_S + sum(surv_val[t_idx:(T_idx - 1)] * dt[t_idx:(T_idx - 1)])

  } else {
    # jeżeli b poza zakresem KM - dopisujemy punkt (b, S_last)
    df.complete <- rbind(df.complete, data.frame(time = b, surv = tail(Surv, 1)))
    dt <- diff(df.complete$time)
    surv_val <- head(df.complete$surv, -1)
```

```
    # całkowanie po całej reszcie przedziału
    integral_S <- integral_S + sum(surv_val[t_idx:length(surv_val)] * dt[t_idx:length(dt)])
  }

  return(integral_S)
}

var.estimator <- function(df, tau) {
  Time <- df$times
  Deltas <- df$deltas
  event_times <- sort(unique(Time[Deltas == 1]))
  result <- 0

  for (i in 1:length(event_times)) {
    t_i <- event_times[i]
    d <- 0
    r <- sum(Time >= t_i)
    for (j in 1:length(Time)) {
      if (Time[j] == t_i & Deltas[j] == 1) {
        d <- d + 1
      }
    }
    if (r > d)
      result <- result + (integrate.S(df, t_i, tau))^2 * d/(r * (r - d))
  }
  return(result)
}

confidence.interval <- function(df, alfa, tau) {
  mu <- integrate.S(df, 0, tau)
  kwantyl <- qnorm(1 - alfa /2)
  std.dev <- sqrt(var.estimator(df, tau))
  T_L <- mu - kwantyl * std.dev
  T_U <- mu + kwantyl * std.dev
  return(data.frame(T_L = T_L, T_U = T_U))
}
```

3.2 Zadanie 2

Zadanie polega na wyznaczeniu realizacji przedziałów ufności, na poziomie ufności 0.95 ($\alpha = 0.05$), dla średniego czasu do progresji choroby w dwóch grupach pacjentek kliniki Mayo Clinic, obejmujących pacjentki chore na raka jajnika w II oraz IIIA stadium choroby.

W przypadku pacjentek w stadium II czasy (w dniach) do progresji choroby były następujące: 28, 89, 175, 195, 309, 377+, 393+, 421+, 447+, 462, 709+, 744+, 770+, 1106+ ,1206+. Symbol „+" oznacza obserwację prawostronnie cenzurowaną.

Dla pacjentek w stadium IIIA czasy (w dniach) wynosiły: 34, 88, 137, 199, 280, 291, 299+, 300+, 309, 351, 358, 369, 369, 370, 375, 382, 392, 429+, 451, 1119+.

Ze względu na fakt, że progresja chorób nowotworowych zależy od wielu czynników, w szczególności indywidualnych, trudno jest ustalić jednoznacznie "sensowne" maksymalne czasy do progresji choroby. Dlatego przedziały ufności wyznaczono dla dwóch wartości parametru τ: $\tau = 1460$ oraz $\tau = 1825$, odpowiadających

odpowiednio 4 i 5 latom (przy przyjęciu uproszczenia: 1 rok = 365 dni). Następnie porównano uzyskane przedziały ufności dla obu grup i obu wartości τ.

```
times.II <- c(28, 89, 175, 195, 309, 377, 393, 421, 447, 462, 709, 744, 770,
              1106, 1206)
deltas.II <- c(1,1,1,1,1,0,0,0,0,1,0,0,0,0,0)
df.II <- data.frame(times = times.II, deltas = deltas.II)

times.III <- c(34, 88, 137, 199, 280, 291, 299, 300, 309, 351, 358, 369, 369,
               370, 375, 382, 392, 429, 451, 1119)
deltas.III <- c(1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1,0,1,0)
df.III <- data.frame(times = times.III, deltas = deltas.III)

tau = 1460
Ci_II_1 <- confidence.interval(df.II, 0.05, tau)
Ci_III_1 <- confidence.interval(df.III, 0.05, tau)

tau = 1825
Ci_II_2 <- confidence.interval(df.II, 0.05, tau)
Ci_III_2 <- confidence.interval(df.III, 0.05, tau)
```

Tabela 3: Realizacje przedziałów przeżycia dla pacjentek z rakiem jajnika typu II i typu IIIA - Mayo Clinic - $\tau = 1460$

Grupa	T_L	T_U
Typ II	578.769	1252.253
Typ IIIA	235.024	584.384

Z Tabeli 3 wynika, że przedział ufności dla pacjentek z rakiem typu II jest szerszy niż dla pacjentek z rakiem typu IIIA. Prawdopodobnie wynika to z faktu, że grupa typu II obejmuje mniej zarejestrowanych obserwacji, w tym mniej obserwacji kompletnych. Dodatkowo, niższe stadium choroby charakteryzuje się większymi wartościami czasu - różnica między minimalną wartością dla typu II a maksymalną dla typu IIIA wynosi zaledwie 5.615. Obserwacja ta odpowiada rzeczywistym danym: wyższy stopień zaawansowania nowotworu wiąże się z szybszą progresją choroby.

Tabela 4: Realizacje przedziałów przeżycia dla pacjentek z rakiem jajnika typu II i typu IIIA - Mayo Clinic - $\tau = 1825$

Grupa	T_L	T_U
Typ II	680.117	1556.460
Typ IIIA	213.670	669.613

Z Tabeli 4 można odczytać odmienne zachowanie przedziałów ufności dla poszczególnych grup. W przypadku pacjentek z nowotworem typu II zarówno dolna, jak i górna granica przedziału wzrosły, przy czym górna granica zwiększyła się wyraźnie bardziej niż dolna. W przypadku raka typu IIIA górna granica również wzrosła, natomiast dolna uległa obniżeniu. Zmiany w grupie typu IIIA są mniejsze niż w grupie typu II, co sugeruje bardziej stabilne zachowanie estymatora funkcji przeżycia w ogonie rozkładu.

Analizując estymacje można zauważyć, że dla raka typu II wartość czasu przeżycia była większa niż jego wariancja, natomiast dla typu IIIA sytuacja była odwrotna. Wspólną cechą obu grup jest wydłużenie przedziałów ufności, co wskazuje na zwiększoną niepewność estymacji przy dłuższym czasie obserwacji.

4 Lista 8

Lista polega na weryfikacji hipotezy o jednakowym rozkładzie czasu do progresji choroby w dwóch badanych grupach pacjentek, przy użyciu testów logrank, Gehana-Breslowa, Tarone'a-Warego i Peto-Peto, porównaniu otrzymanych wartości poziomów krytycznych oraz na naszkicowaniu wykresów estymatorów Kaplana-Meiera funkcji przeżycia dla czasu do progresji choroby w obu grupach wraz z funkcjami wagowymi użytymi w statystykach testowych.

4.1 Zadanie 1

W zadaniu dokonano wyliczenia wartości p-value dla hipotezy o jednakowym rozkładzie czasu do progresji choroby w dwóch badanych grupach pacjentek, z podziałem na sposób dobierania wagi.

Na początek łączymy poszczególne próby $\mathbb{X}_1, \ldots, \mathbb{X}_k$, gdzie $\mathbb{X}_j = (X_{j1}, \ldots, X_{jn_j})$, $j \in \{1, \ldots, k\}$. Z tak połączonych grup wyznaczamy (unikalne) czasy zdarzeń, których liczbę oznaczamy przez D, i porządkujemy je rosnąco: $t_1 < t_2 < \cdots < t_D$.

Statystyka testowa opiera się na sumie ważonych różnic między oszacowanymi hazardami w momentach t_i. Oszacowania te wyznaczane są za pomocą estymatora Nelsona–Aalena funkcji hazardu:

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(t_i) \left(\frac{d_{ij}}{r_{ij}} - \frac{d_i}{r_i} \right), \quad \text{gdzie:}$$

- d_{ij} – liczba zdarzeń w czasie t_i z próby $\mathbb{X}_j$,
- r_{ij} – liczba jednostek w stanie ryzyka w czasie t_i w j-tej grupie,
- $d_i = \sum_{j=1}^{k} d_{ij}$,
- $r_i = \sum_{j=1}^{k} r_{ij}$.

W zadaniu przyjęto funkcję wagi jako $W_j(t_i) = r_{ij} W(t_i)$, wtedy statystyka $Z_j(\tau)$ ma postać:

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_i) \left(d_{ij} - r_{ij} \frac{d_i}{r_i} \right).$$

Wariancję takiej statystyki estymujemy przez:

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W^2(t_i) d_i \frac{r_{ij}}{r_i} \left(1 - \frac{r_{ij}}{r_i} \right) \left(\frac{r_i - d_i}{r_i - 1} \right),$$

kowariancję natomiast szacujemy jako:

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W^2(t_i) d_i \frac{r_{ij} r_{ig}}{r_i^2} \left(\frac{r_i - d_i}{r_i - 1} \right).$$

Statystyka testowa do weryfikacji rozpatrywanych hipotez konstruowana jest w oparciu o dowolne $k-1$ wybranych statystyk $Z_j(\tau)$ i ma postać:

$$Z^2 = (Z_1(\tau), \ldots, Z_{k-1}(\tau)) \, \Sigma^{-1} \, (Z_1(\tau), \ldots, Z_{k-1}(\tau))^T,$$

gdzie macierz Σ jest utworzona z odpowiednich elementów $\hat{\sigma}_{jj}$ oraz $\hat{\sigma}_{jg}$.

Udowodniono, że jeżeli hipoteza H_0 jest spełniona, to rozkład statystyki Z^2 dąży według rozkładu do rozkładu chi-kwadrat z $k-1$ stopniami swobody, gdy $n_1 \to \infty, \ldots, n_{k-1} \to \infty$. Korzystając z tego faktu, można obliczyć wartość poziomu krytycznego:

$$p = 1 - F_{\chi^2_{k-1}}(z^2),$$

gdzie $F_{\chi^2_{k-1}}$ jest dystrybuantą rozkładu chi-kwadrat z $k-1$ stopniami swobody, a z^2 realizacją statystyki Z^2.

W zadaniu dokonano czterech testów ze względu na dobraną postać funkcji $W(t_i)$:

- test logrank: $W(t_i) \equiv 1$,
- test Gehana–Breslowa: $W(t_i) = r_i$,
- test Peto–Peto: $W(t_i) = \prod_{t_j \leq t_i} \left(1 - \frac{d_j}{r_j+1}\right)$,
- test Tarone'a–Ware'a: $W(t_i) = \sqrt{r_i}$.

Tabela 5: Wartości p-value dla testów czasu do progresji choroby weryfikujących hipotezę o jednakowym rozkładzie w grupach pacjentek rakowych kliniki Mayo Clinic

Test	p-value
logrank	0.019
Gehan-Breslow	0.134
Peto-Peto	0.098
Tarone-Ware	0.055

Z Tabeli 5 wynika, że na ustalonym poziomie istotności $\alpha = 0.05$ hipoteza o jednakowym rozkładzie czasu do progresji powinna zostać odrzucona jedynie w przypadku testu logrank. Wartości p-value dla poszczególnych funkcji wag różnią się między sobą, co wynika bezpośrednio z ich konstrukcji:

- **logrank** traktuje wszystkie zdarzenia równoważnie, wykrywając globalne różnice między grupami, nawet jeśli ujawniają się one późno,
- **Gehan-Breslow** przypisuje większą wagę wcześniejszym czasom - jeżeli różnice między grupami pojawiają się późno, test staje się na nie mniej wrażliwy,
- **Tarone-Ware** jest kompromisem między logrank a Gehan-Breslow - zmniejsza różnice między wagami, nakładając na nie pierwiastek,
- **Peto-Peto** dla poszczególnych czasów uwzględnia nie tylko aktualne zdarzenie, ale również wszystkie wcześniejsze oraz informację o liczbie zdarzeń w połączonych próbach, które wystąpiły w tym samym czasie.

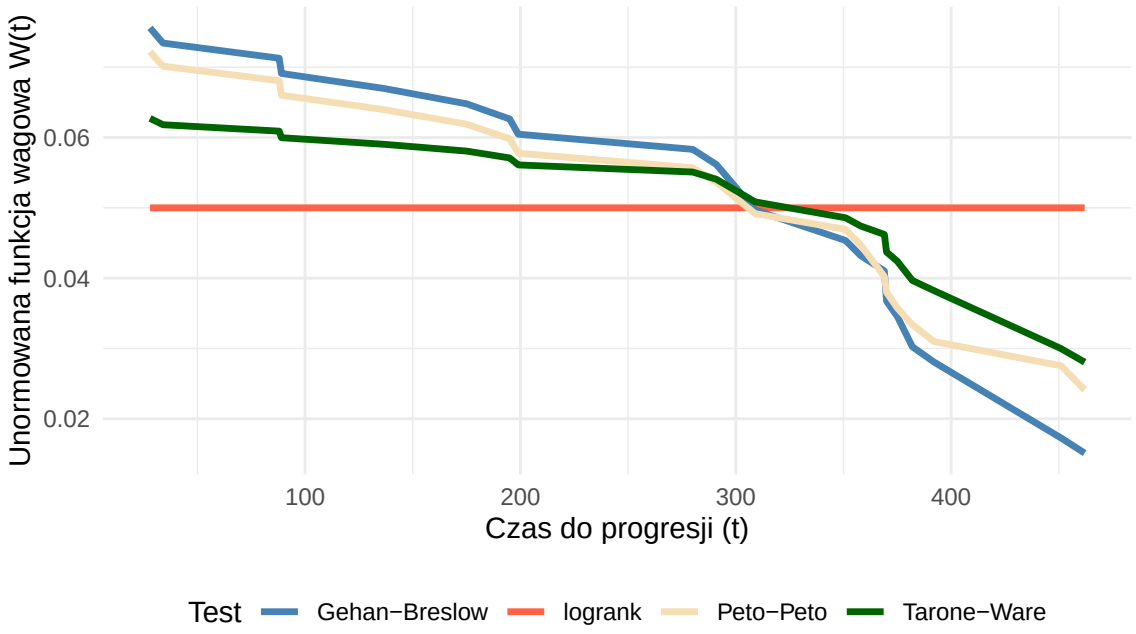
Podsumowując, istotność wykazana przez test logrank wskazuje na globalną różnicę w przebiegu czasu do zdarzenia między grupami, natomiast brak istotności w pozostałych testach odzwierciedla fakt, że lokalne lub czasowo rozłożone różnice są mniej wyraźne.

4.2 Zadanie 2

Zadanie polega na naszkicowaniu wykresów estymatorów Kaplana–Meiera funkcji przeżycia dla dwóch grup oraz unormowanych funkcji wagowych testów logrank, Gehan-Breslow, Tarone-Ware i Peto-Peto. Na podstawie wykresów należy przeanalizować, jak kształt funkcji przeżycia i przypisane wagi wpływają na wartości p-value oraz wnioski dotyczące istotności różnic między grupami.

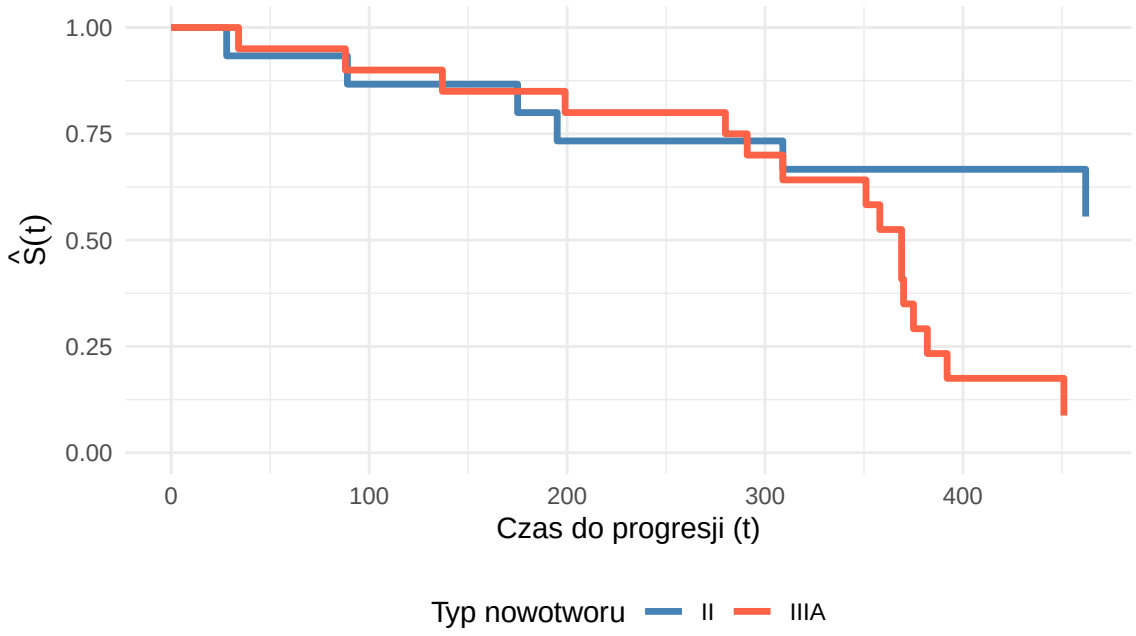
Poniżej przedstawiono blok kodu w R, który oblicza dla każdego zaobserwowanego czasu zdarzenia liczby jednostek w stanie ryzyka i liczby zdarzeń, a następnie na ich podstawie generuje unormowane funkcje wag.

```
# Uzupełnienie danych o wartości r_i i d_i
df.cancer$r <- 0
df.cancer$d <- 0
for(i in 1:nrow(df.cancer)){
  df.cancer[i, "r"] <- sum(df.cancer$times >= df.cancer[i, "times"])
  df.cancer[i, "d"] <- sum(df.cancer$times == df.cancer[i, "times"] &
                           df.cancer$deltas==1)
}

# Unikalne zdarzenia
df.cancer_unique <- df.cancer[!duplicated(df.cancer$times) &
                           df.cancer$deltas==1,]

# Wagi testu logrank
df.cancer_unique$lg <- 1

# Wagi testu Gehana-Breslowa
df.cancer_unique$GB <- df.cancer_unique$r

# Wagi testu Peto-Peto
df.cancer_unique$PP <- 0
for(i in 1:nrow(df.cancer_unique)){
  czas <- df.cancer_unique[i, "times"]
  iloczyn <- 1
  for(j in 1:nrow(df.cancer_unique)){
    if(df.cancer_unique[j, "times"] <= czas){
      dj <- df.cancer_unique[j, "d"]
      rj <- df.cancer_unique[j, "r"]
      iloczyn <- iloczyn * (1 - dj/(rj + 1))
    }
  }
  df.cancer_unique[i, "PP"] <- iloczyn
}

# Wagi testu Tarone'a-Were'a
df.cancer_unique$TW <- sqrt(df.cancer_unique$r)

# Unormowanie wag
df.cancer_unique$lg <- df.cancer_unique$lg/sum(df.cancer_unique$lg)
df.cancer_unique$GB <- df.cancer_unique$GB/sum(df.cancer_unique$GB)
df.cancer_unique$PP <- df.cancer_unique$PP/sum(df.cancer_unique$PP)
df.cancer_unique$TW <- df.cancer_unique$TW/sum(df.cancer_unique$TW)
```



Wykres 7: Wagi testów nieparametrycznych dla porównania rozkładów czasów przeżycia - Mayo Clinic

Na Wykresie 7 widać, jakie unormowane wagi przypisuje każdy z testów dla poszczególnych czasów. Dla $t \approx 320$ wagi testów Gehan-Breslow, Peto-Peto i Tarone-Ware przecinają krzywą wag testu logrank, co oznacza, że dla zdarzeń o czasie większym niż 320 przypisywane jest mniejsze znaczenie niż przy teście logrank.



Wykres 8: Estymator Kaplana-Meiera dla danych dotyczących pacjentek z rakiem jajnika - Mayo Clinic

Na Wykresie 8. można zauważyć, że estymatory przeżycia Kaplana-Meiera dla obu zbiorów danych pacjentek rakowych są dość zbliżone do czasu $t \approx 320$. Po tym punkcie funkcje przeżycia gwałtownie spadają, co oznacza, że w tym okresie występuje progresja choroby dla wielu pacjentek. To skutkuje spadkiem wag testów Gehan-Breslow, Peto-Peto i Tarone-Ware w późniejszych momentach czasu, co oznacza, że późne zdarzenia mają mniejszy wpływ na wartość statystyki testowej w porównaniu do testu logrank. W konsekwencji główny wkład do obliczenia wartości p pochodzi ze zdarzeń występujących wcześniej, zanim funkcje przeżycia drastycznie spadną.

Ze względu na spadek wag testów Gehan-Breslow, Peto-Peto i Tarone-Ware dla późnych czasów, istotność statystyczna opiera się głównie na wcześniejszych zdarzeniach. W tym przypadku najbardziej reprezentatywne jest p-value z testu logrank, który równomiernie uwzględnia wszystkie zdarzenia, co widać na Wykresie 8. estymatorów Kaplana-Meiera.

5 Zadania dodatkowe

5.1 Zadanie 1

```r
cenzurowanie_losowe <- function(theta, data) {
  censor_data <- rexp(length(data), rate=theta)
  deltas <- as.numeric(data <= censor_data)
  censored_data <- ifelse(data <= censor_data, data, censor_data)
  list(times = censored_data, deltas = deltas)
}

GE_cenzurowanie_losowe <- function(theta, alpha, lambd, n) {
  data <- EW_generator(alpha = 1, beta = lambd, gamm = alpha, size = n)
  cenzurowanie_losowe(theta, data)
}
```

5.2 Zadanie 2

Estymatory Kaplana-Meiera oraz Fleminga–Harringtona są funkcjami schodnowymi. W stosunkowo małych próbach często obserwuje się szerokie przedziały czasu w których estymator jest stały - praktycznie i interpretacyjnie bywa to niepożądane i statystykom trudno jest wytłumaczyć praktykom na przykładzie Wykresu 1. grupy leku A dlaczego prawdopodobieńsgtwo przeżycia do chwili $t = 0.7$ jest równe prawdopodobieństwu przeżycia do chwili $t = 0.8$.

Rossa i Zieliński [2] zaproponowali lokalne „wygładzenie" estymatora Kaplana–Meiera poprzez dopasowanie funkcji gęstości rozkładu Weibulla w niewielkim otoczeniu każdego punktu skoku.

Zasada działania estymatora

Krok 1. Konstrukcja wartości środkowych

1. Niech $T_0 = 0$ oraz $KM(T_0) = 1$, a

   $$T_{(1)} < T_{(2)} < \cdots < T_{(N-1)}$$

   będą uporządkowanymi czasami obserwacji, gdzie $T_{(1)}, \ldots, T_{(N-1)}$ oznaczają momenty skoków klasycznego estymatora Kaplana–Meiera, natomiast $T_{(N)}$ jest ostatnim czasem w próbie (zdarzeniem lub cenzurowaniem). Wartości $KM\left(T_{(j)}\right)$ oznaczają wartość estymatora Kaplana–Meiera w chwili $T_{(j)}$, tj. tuż po wystąpieniu skoku.

2. Dla każdego $j = 1, \ldots, N-1$ wartość środkową definiuje się jako średnią arytmetyczną dwóch kolejnych wartości estymatora:

   $$KM'_j = \frac{KM\left(T_{(j-1)}\right) + KM\left(T_{(j)}\right)}{2}.$$

3. W przypadku ostatniego czasu skoku $T_{(N)}$ wartość KM'_N definiuje się jako:

   $$KM'_N = \begin{cases} \dfrac{KM\left(T_{(N)}\right)}{2}, & \delta_N = 1, \\ KM\left(T_{(N)}\right), & \delta_N = 0. \end{cases}$$

Krok 2. Obliczanie wag

Dla dowolnego $t \geq 0$ definiujemy wektor wag

$$\mathbf{w}(t) = (w_1(t), \ldots, w_N(t)) .$$

Wagi zależą od parametru m — liczby "sąsiadów" (liczb całkowitych) — oraz od położenia t względem przedziałów między momentami skoków estymatora. Wagi te będą użyte w następnych krokach przy regresji liniowej.

Rozróżniamy trzy przypadki dla różych m:

A) Gdy $m = 2$ wagi są zawsze "punktowe", a sam estymator przyjmuje odmienną, uproszczoną postać. Szczegóły przedstawiono w dalszej części.

B) Gdy m jest nieparzyste tzn. $m = 2k + 1$ dla pewnego $k \in \mathbb{N}$. Wagi definiujemy następująco:

Zdefiniujmy najpierw indeks j opisujący, w którym przedziale znajduje się t:

$$j = \begin{cases} 0, & t \leq T_{(1)}, \\ N, & t > T_{(N)}, \\ i \in 1, \ldots, N-1, & T_{(i)} < t \leq T_{(i+1)} \end{cases}$$

1. Jeżeli $j \leq k$ (t blisko początku) - krawedź lewa:

$$w_1(t) = w_2(t) = \cdots = w_m(t) = 1,$$

oraz

$$w_{m+1}(t) = \cdots = w_N(t) = 0.$$

2. Jeżeli $j \geq N - k$ (t blisko końca) - krawedź prawa:

$$w_{N-m+1}(t) = w_{N-m+2}(t) = \cdots = w_N(t) = 1,$$

oraz

$$w_1(t) = \cdots = w_{N-m}(t) = 0.$$

3. Jeżeli $k < j < N-k$ (t jest dostatecznie daleko od krawedzi i otoczenie o rozmiarze m mieści się w zakresie indeksów) - przypadek środkowy:

1. Blok jednostkowy (dokładnie m indeksów) Dla indeksów $i \in j-k, j-k+1, \ldots, j+k$ przyjmujemy $w_i(t) = 1$.

2. Dwa brzegowe indeksy sąsiednie (ewentualne wagi ułamkowe) Jeśli istnieje indeks lewy ($j - k \geq 1$)

$$i_{\text{lewy}} = j - k,$$

to jego waga zależy liniowo od położenia t w przedziale $(T_{(j)}, T_{(j+1)}]$:

$$w_{i_{\text{lewy}}}(t) = \frac{T_{(j+1)} - t}{T_{(j+1)} - T_{(j)}}.$$

Jeśli istnieje indeks prawy ($j + k + 1 \leq N$))

$$i_{\text{prawy}} = j + k + 1,$$

to jego waga wynosi

$$w_{i_{\text{prawy}}}(t) = \frac{t - T_{(j)}}{T_{(j+1)} - T_{(j)}}.$$

3. Pozostałe wagi

   Wszystkie pozostałe składowe wektora wag przyjmują wartość 0.

Gdy m jest parzyste tzn. $m = 2k$ dla pewnego $k \in \mathbb{N}_+$. Wagi definiujemy następująco:

Zdefiniujmy najpierw indeks j opisujący, w którym przedziale znajduje się t:

$$j = \begin{cases} 0, & t \leq T_{(1)}/2, \\ N, & t > (T_{(N-1)} + T_{(N)})/2, \\ i \in 1, \ldots, N-1, & \left(T_{(i-1)} + T_{(i)}\right)/2 < t \leq \left(T_{(i)} + T_{(i+1)}\right)/2 \end{cases}$$

1. Jeżeli $j \leq k$ (t blisko początku) - krawedź lewa:

$$w_1(t) = w_2(t) = \cdots = w_m(t) = 1,$$

oraz

$$w_{m+1}(t) = \cdots = w_N(t) = 0.$$

2. Jeżeli $j \geq N - k + 1$ (t blisko końca) - krawedź prawa:

$$w_{N-m+1}(t) = w_{N-m+2}(t) = \cdots = w_N(t) = 1,$$

oraz

$$w_1(t) = \cdots = w_{N-m}(t) = 0.$$

3. Jeżeli $k < j < N - k + 1$ (t jest dostatecznie daleko od krawedzi i otoczenie o rozmiarze m mieści się w zakresie indeksów) - przypadek środkowy:

4. Blok jednostkowy (dokładnie m indeksów) Dla indeksów $i \in j - k + 1, j - k + 2, \ldots, j + k$ przyjmujemy $w_i(t) = 1$.

5. Dwa brzegowe indeksy sąsiednie (ewentualne wagi ułamkowe) Jeśli istnieje indeks lewy ($j - k \geq 1$)

   $$i_{\text{lewy}} = j - k,$$

   to jego waga zależy liniowo od położenia t w przedziale $(T_{(j)}, T_{(j+1)}]$:

   $$w_{i_{\text{lewy}}}(t) = \frac{\frac{1}{2}\left(T_{(j)} + T_{(j+1)}\right) - t}{\frac{1}{2}\left(T_{(j+1)} - T_{(j-1)}\right)}.$$

   Jeśli istnieje indeks prawy ($j + k \leq N$))

   $$i_{\text{prawy}} = j + k,$$

   to jego waga wynosi

   $$w_{i_{\text{prawy}}}(t) = \frac{t - \frac{1}{2}\left(T_{(j)} + T_{(j+1)}\right)}{\frac{1}{2}\left(T_{(j+1)} - T_{(j-1)}\right)}.$$

6. Pozostałe wagi

   Wszystkie pozostałe składowe wektora wag przyjmują wartość 0.

Krok 3. Interpolacja liniowa oraz ważona regresja liniowa

W pierwszej kolejności obliczamy log-czasy zdarzeń oraz ich log-log przeżycia, które następnie stanowią dane wejściowe do lokalnej regresji liniowej budującej estymator Rossy-Zielińskiego.

$$X_j = \log T_{(j)}, \qquad Y_j = \log\left(-\log(KM'_j)\right),$$

W przypadku $m = 2$ i w ustalonej chwili t obliczamy log-t:

$$X = \log t$$

Estymator $S_2(t)$ jest liniową transformacją log-log:

$$Y(t) = \begin{cases} Y_1 + \dfrac{X_2 - X_1}{Y_2 - Y_1}(X - X_1), & 0 < t \leq T_{(1)}, \\ Y_j + \dfrac{X_{j+1} - X_j}{Y_{j+1} - Y_j}(X - X_{j-1}), & T_{(j)} < t \leq T_{(j+1)}, \\ Y_{N-1} + \dfrac{X_N - X_{N-1}}{Y_N - Y_{N-1}}(X - X_{N-1}), & t > T_{(N)} \text{ i } \delta_N = 1, \\ \text{nieokreślony}, & t > T_{(N)} \text{ i } \delta_N = 0. \end{cases}$$

W przypadku $m > 2$ i w ustalonej chwili t, dysponujemy wektorem wag:

$$\mathbf{w}(t) = (w_1(t), \ldots, w_N(t)),$$

Dopasowanie lokalnego modelu Weibulla odbywa się przez zastosowanie ważonej regresji liniowej w przestrzeni transformowanej:

$$Y_i = aX_i + b + \varepsilon_i,$$

gdzie parametry $a = a(t)$ i $b = b(t)$ wyznacza się jako rozwiązanie problemu:

$$\min_{a,b} \sum_{i=1}^{N} w_i(t) \left[Y_i - (aX_i + b)\right]^2.$$

Wynikiem są lokalne estymatory:

$$\widehat{a}(t), \qquad \widehat{b}(t),$$

które interpretujemy jako lokalne parametry kształtu i skali rozkładu Weibulla dopasowanego w "otoczeniu" punktu "t" przy użyciu wag opisanych wcześniej.

Krok 4. Transformacja odwrotna

W przypadku $m = 2$ wartość estymatora Rossy-Zielińskiego definuje się jako:

$$\widehat{S}_2(t) = \exp\left(-\exp(Y(t))\right).$$

W przypadku $m > 2$ rekonstruujemy lokalną funkcję przeżycia, korzystając z postaci funkcji przeżycia rozkładu Weibulla:

$$S(t) = \exp\left(-\exp(a)t^b\right).$$

Estymator Rossy–Zielińskiego definiuje się jako:

$$\widehat{S}_m(t) = \exp\left(-\exp\left(\widehat{a}(t)\right) t^{\widehat{b}(t)}\right).$$

Otrzymana funkcja jest lokalnie wygładzona, a stopień wygładzenia kontroluje parametr m, który określa szerokość okna oraz sposób formowania wag.

Poniżej przedstawiono blok kodu w języku R implementujący estymator Rossy–Zielińskiego wraz z przykładowym generowaniem wykresów funkcji przeżycia.

```r
rossa.zielinski.estimator <- function(df, m = 2) {
  # Funkcja pomocnicza do liczenia wag (dla m > 2)
  compute_weights <- function(Time, t, m) {
    N <- length(Time)
    w <- rep(0, N)

    # m = 2k + 1 - nieparzyste
    if (m %% 2 == 1) {
      k <- (m - 1) / 2
      if (t <= Time[1]) {
        j <- 0
      } else if (t > Time[N]) {
        j <- N
      } else {
        j <- max(which(Time < t))
      }
      # generowanie wag
      if (j <= k) {
        w[1:m] <- 1
      }
      else if (j >= N - k) {
        w[(N - m + 1):N] <- 1
      }
      else {
        idx_unit <- (j - k + 1):(j + k)
        w[idx_unit] <- 1
        # przypadki brzegowe
        if ((j - k) >= 1) {
          w[j - k] <- (Time[j + 1] - t) / (Time[j + 1] - Time[j])
        }
        if ((j + k + 1) <= N) {
          w[j + k + 1] <- (t - Time[j]) / (Time[j + 1] - Time[j])
        }
      }
      # m = 2k - parzyste
    } else {
      k <- m / 2
      if (t <= (Time[1]/2)) {
        j <- 0
      } else if (t > (Time[N-1] + Time[N])/2) {
        j <- N
      } else {
        j <- which((Time[-1] + Time[-N])/2 >= t)[1]
      }
```

```
    # generowanie wag
    if (j <= k) {
      w[1:m] <- 1
    } else if (j >= N - k + 1) {
      w[(N - m + 1):N] <- 1
    } else {
      idx_unit <- (j - k + 1):(j + k - 1)
      w[idx_unit] <- 1

      # przypadki brzegowe
      if ((j - k) >= 1) {
        w[j - k] <- (0.5*(Time[j] + Time[j + 1]) - t) /
          0.5 * (Time[j + 1] - Time[j - 1])
      }
      if ((j + k) <= N) {
        w[j + k] <- (t - 0.5*(Time[j - 1] + Time[j])) /
          0.5 * (Time[j + 1] - Time[j - 1])
      }
    }
  }
  return(w)
}

# wartości środkowe KM'
fit <- survfit(Surv(df$times, df$deltas) ~ 1)
Time <- fit$time[fit$n.event > 0]
Time <- c(Time, tail(fit$time, 1))
KM <- fit$surv[fit$n.event > 0]
KM <- c(KM, tail(fit$surv, 1))
N <- length(Time)

KM0 <- numeric(N)
for (i in 1:(N-1)) {
  if (i == 1) {
    KM0[i] <- (1 + KM[i]) / 2
  } else {
    KM0[i] <- (KM[i-1] + KM[i]) / 2
  }
}
last_delta <- fit$n.event[which(fit$time == Time[N])]
if (last_delta == 1) {
  KM0[N] <- KM[N] / 2
} else {
  KM0[N] <- KM[N]
}

X <- log(Time)
Y <- log(-log(KM0))

# Funkcja estymatora
estimator <- function(x.eval) {
  S <- numeric(length(x.eval))
```

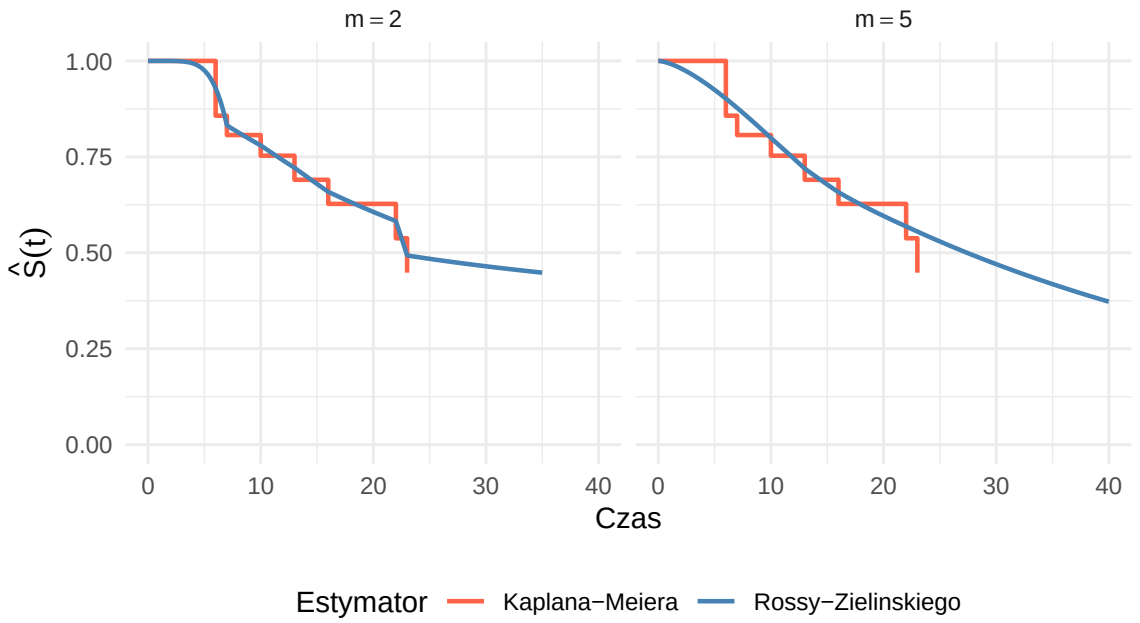
```
    for (i in seq_along(x.eval)) {
      x <- x.eval[i]
      Xx <- log(x)

      # Estymator dla m = 2
      if (m == 2) {
        if (x <= Time[1]) {
          Yx <- Y[1] + (Y[2]-Y[1]) / (X[2]-X[1]) * (Xx - X[1])
        } else if (x > Time[N]) {
          last_delta <- fit$n.event[which(fit$time == Time[N])]
          if (last_delta == 1) {
            Yx <- Y[N-1] + (Y[N]-Y[N-1]) / (X[N]-X[N-1]) * (Xx - X[N-1])
          } else {
            Yx <- NA
          }
        } else {
          j <- max(which(Time < x))
          Yx <- Y[j] + (Y[j+1]-Y[j]) / (X[j+1]-X[j]) * (Xx - X[j])
        }
        if (is.na(Yx)) {
          S[i] <- NA
        } else {
          S[i] <- exp(-exp(Yx))
        }

      } else {
        # Klasyczny przypadek m > 2 z wagami i regresją
        w <- compute_weights(Time, x, m)
        reg <- lm(Y ~ X, weights = w)
        a <- coef(reg)[1]
        b <- coef(reg)[2]
        S[i] <- exp(-exp(a) * x^b)
        S[i] <- max(min(S[i],1),0)
      }
    }
    return(S)
  }
  return(estimator)
}

rz.estimator.plot <- function(df, m = 2, to = 3) {
  estimator <- rossa.zielinski.estimator(df, m)

  time <- seq(0, to, length.out = 1000)
  surv <- estimator(time)

  plot.df <- data.frame(time = time, surv = surv)

  ggplot(plot.df, aes(x = time, y = surv)) +
    geom_line(color = "tomato", linewidth = 1) +
    scale_y_continuous(limits = c(0, 1)) +
    labs(title = "Estymator Rossa-Zieliński",
         x = "Czas do zdarzenia (t)",
```

```
        y = "Funkcja przeżycia S(t)") +
    theme_minimal()
}
```

Porównamy graficznie estmator Kaplana-Maiera z jego ulepszoną wersją estymatorem Rossy-Zielińskiego.

W tym celu sporządzono wykresy dla danych klinicznych pacjentów przedstawionych w artykule Rossy-Zielińskiego, a także dla danych z zadania 3 z listy 2 oraz z zadania 2 z listy 7.

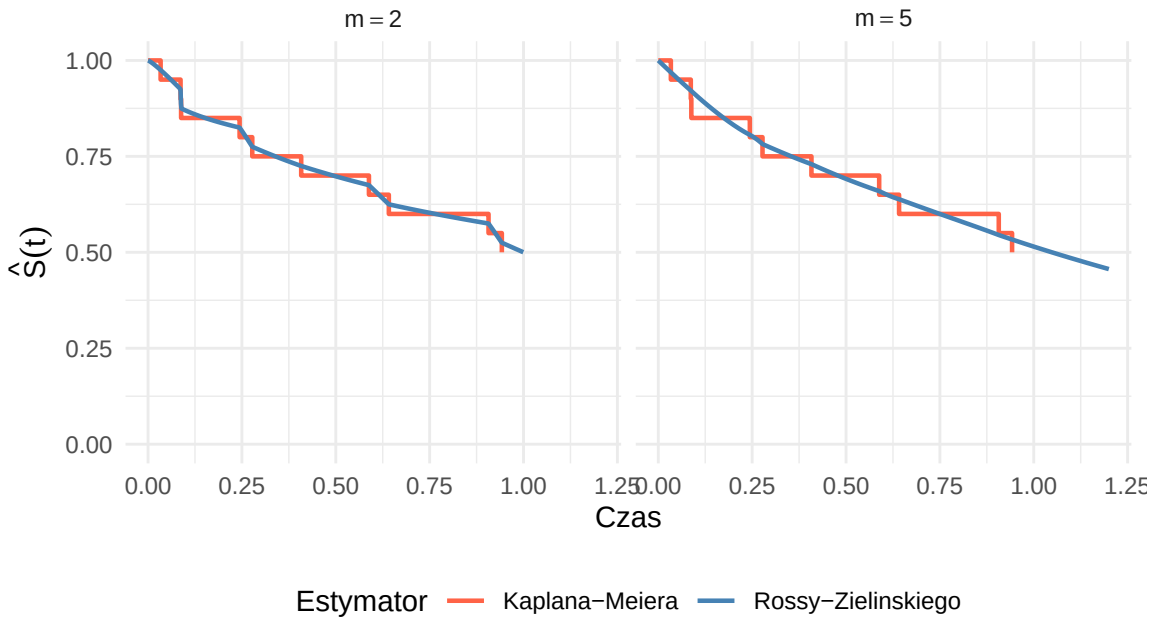
```
## Warning: Removed 125 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Wykres 9: Porównanie wstymatora Kaplana-Meiera oraz Rossy-Zielińskiego dla zbioru danych klinicznych pacjentów

Z Wykresu 9. można odczytać, że generowane wykresy są bardzo podobne do tych przedstawionych w artykule. Można przypuszczać, że opracowany algorytm działa poprawnie.

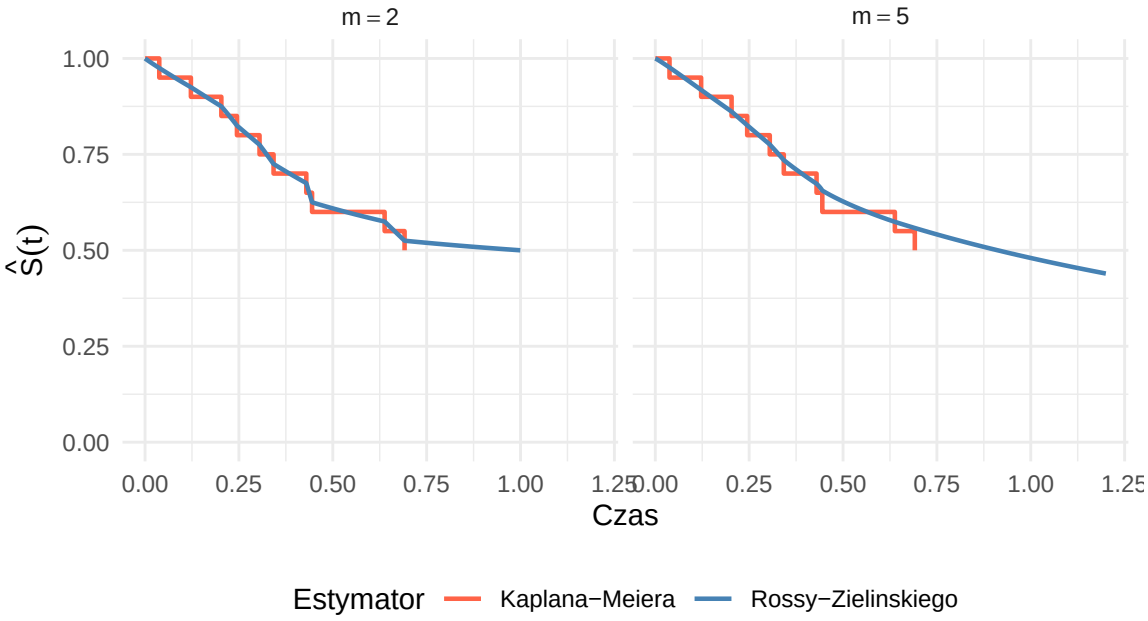
```
## Warning: Removed 167 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Wykres 10: Porównanie wstymatora Kaplana-Meiera oraz Rossy-Zielińskiego dla zbioru danych dotyczących leków - lek A

Wykres 10. pokazuje, w jaki sposób estymator wygładza klasyczny estymator Kaplana–Meiera dla danych dotyczących leków - lek A.

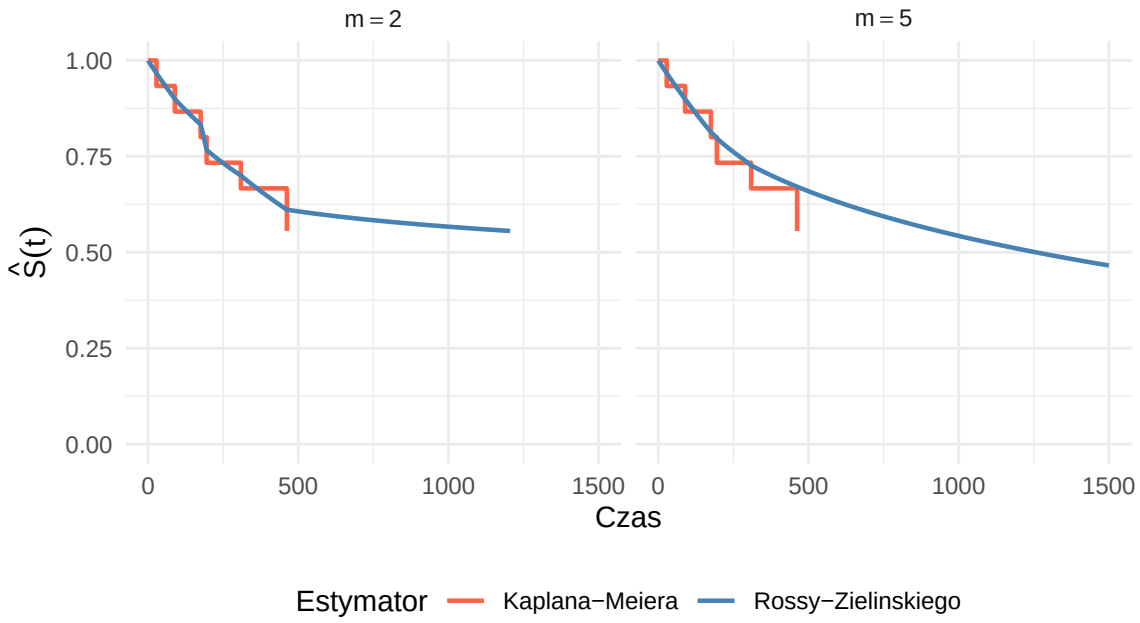
```
## Warning: Removed 167 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Wykres 11: Porównanie wstymatora Kaplana-Meiera oraz Rossy-Zielińskiego dla zbioru danych dotyczących leków - lek B

Wykres 11. pokazuje, w jaki sposób estymator wygładza klasyczny estymator Kaplana–Meiera dla danych dotyczących leków - lek B.

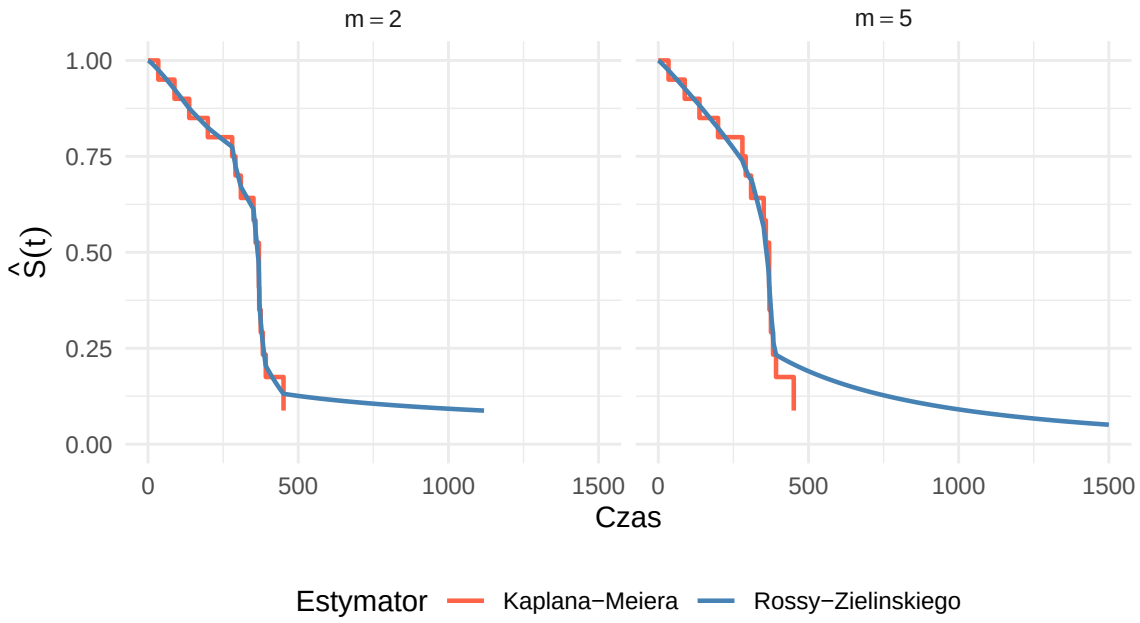
```
## Warning: Removed 196 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Wykres 12: Porównanie wstymatora Kaplana-Meiera oraz Rossy-Zielińskiego dla zbioru danych dotyczących pacjentów z rakiem jajnka - typ II

Wykres 12. pokazuje, w jaki sposób estymator wygładza klasyczny estymator Kaplana–Meiera dla danych dotyczących pacjentów z rakiem jajnika typu II.

```
## Warning: Removed 254 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Wykres 13: Porównanie wstymatora Kaplana-Meiera oraz Rossy-Zielińskiego dla zbioru danych dotyczących pacjentów z rakiem jajnka - typ IIIA

Wykres 13. pokazuje, w jaki sposób estymator wygładza klasyczny estymator Kaplana–Meiera dla danych dotyczących pacjentów z rakiem jajnika typu IIIA. Tym razem czasy są znacznie krótsze, a estymator dla wartości powyżej ostatniej obserwacji "odbija" w prawo. Można zauważyć, że dla $m = 5$ estymator "odbija" nieco wcześniej.

Podsumowując, estymator Rossy-Zielińskiego sprawnie wygładza funkcję przeżycia estymowaną przez Kaplana–Meiera, dobrze dopasowując ją w okolicach czasów zdarzeń, jednocześnie eliminując nagłe skoki i fluktuacje wynikające z małych prób. Dzięki temu estymator daje bardziej stabilny i ciągły obraz ryzyka przeżycia, ułatwiając interpretację oraz porównania między grupami pacjentów.

6 Bibliografia

[1] R. D. Gupta i D. Kundu, *Generalized Exponential Distributions*, Australian & New Zealand Journal of Statistics, 41(2):173–188, 1999, PDF.

[2] Agnieszka Rossa i Ryszard Zieliński, *A Simple Improvement of the Kaplan-Meier Estimator*, Communications in Statistics - Theory and Methods, 31(1):147-158, 2002, DOI: 10.1081/STA-120002440.