Analiza Przeżycia

Raport 2

Filip Michewicz 282239
Wiktor Niedźwiedzki 258882

23 listopada 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 $t_{(i)}$.

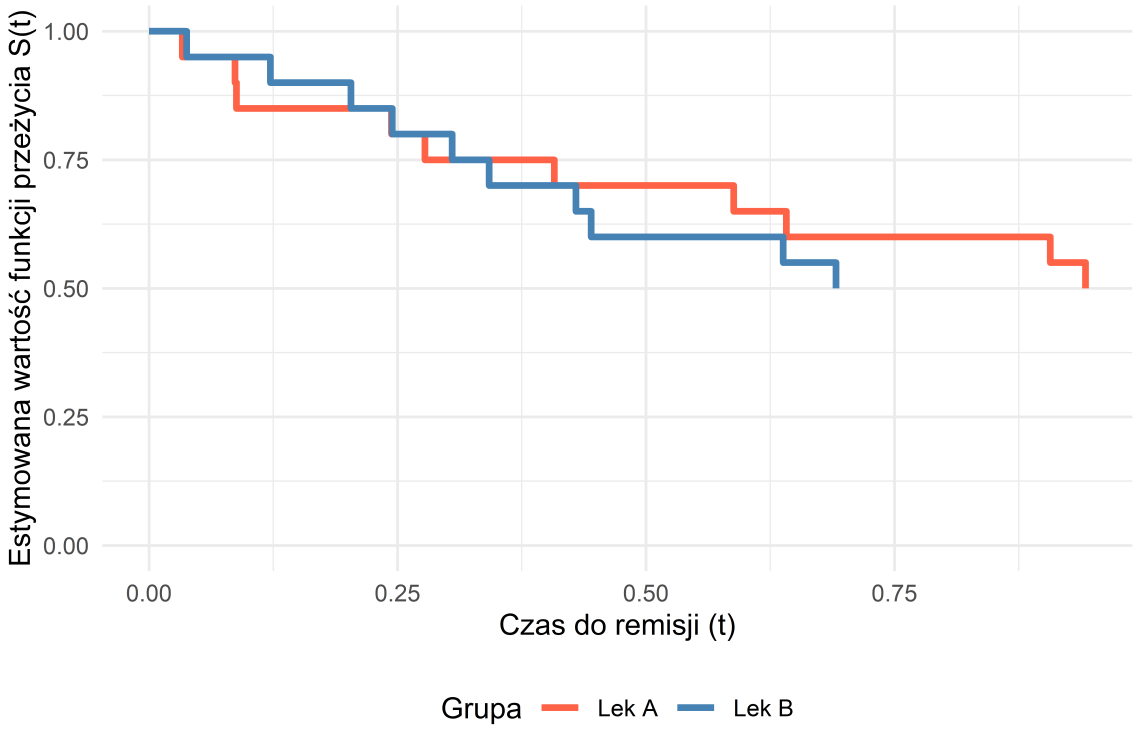
Estymator Kaplana–Meiera nie został zdefiniowany ręcznie w tym raporcie, lecz jego implementacja znajduje się w bibliotece `survival` w pakiecie R. Funkcja `survfit` domyślnie dopasowuje dane w formacie `Surv`, czyli takim, w którym określone są czasy obserwacji oraz wskaźniki cenzurowania, i domyślnie 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
event_idx <- which(fit.KM.A$n.event > 0)
plot.A <- data.frame(time = c(0, fit.KM.A$time[event_idx]),
                     surv = c(1, fit.KM.A$surv[event_idx]),
                     group = "A")

event_idx <- which(fit.KM.B$n.event > 0)
plot.B <- data.frame(time = c(0, fit.KM.B$time[event_idx]),
                     surv = c(1, fit.KM.B$surv[event_idx]),
                     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 niecenzurowane, a po ostatnim pełnym (niecenzurowanym) czasie występują obserwacje cenzurowane, przez co estymator nie jest określony dla $t > t_{(n)}$. Do czasu $t \leq t_{(n)}$ estymator Kaplana–Meiera jest dobrze określony; w naszym przypadku ostatnia obserwacja jest cenzurowana, więc estymator nie jest określony dla $t > t^+$, gdzie t^+ odpowiada czasowi ostatniej obserwacji kompletnej (niecenzurowanej).

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:

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

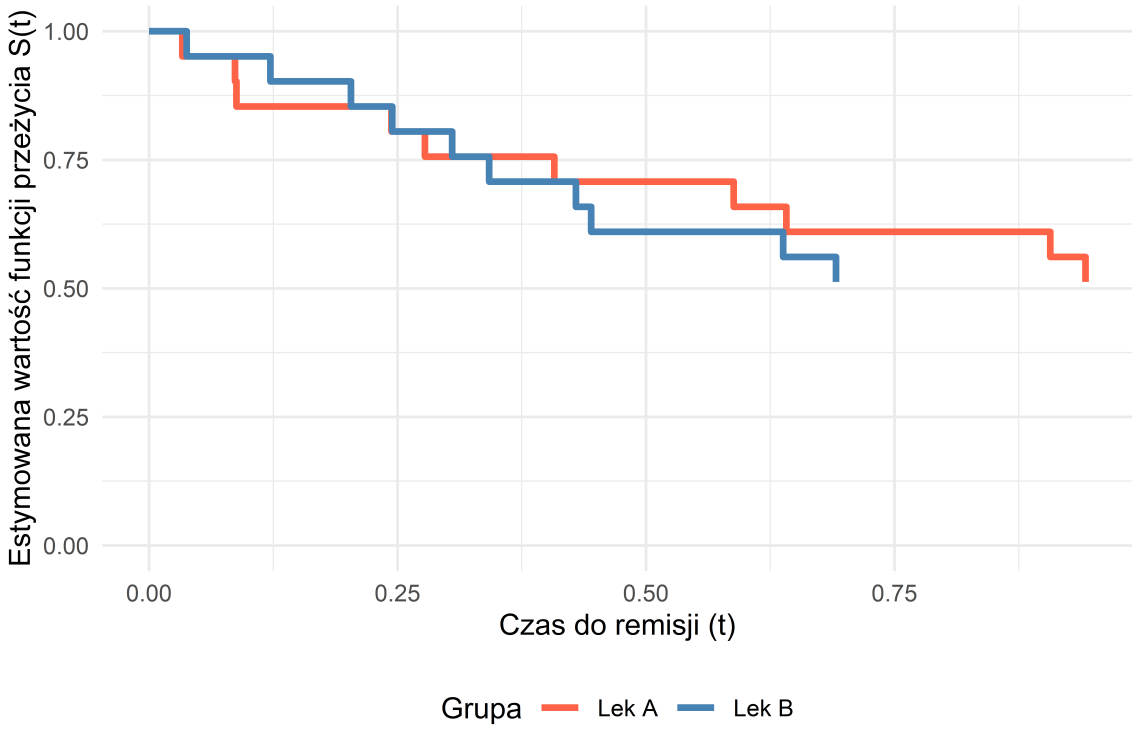
gdzie d_j to liczba zdarzeń w chwili $t_{(j)}$, a r_j - liczba obserwacji pozostających w stanie ryzyka w czasie $t_{(j)}$.

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
event_idx <- which(fit.FH.A$n.event > 0)
plot.A <- data.frame(time = c(0, fit.FH.A$time[event_idx]),
                     surv = c(1, fit.FH.A$surv[event_idx]),
                     group = "A")

event_idx <- which(fit.KM.B$n.event > 0)
plot.B <- data.frame(time = c(0, fit.FH.B$time[event_idx]),
                     surv = c(1, fit.FH.B$surv[event_idx]),
                     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 test, np. 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^+ (czasie ostatniej obserwacji kompletnej) była równa $S(t^+)$.

Po krótkich obliczeniach przeprowadzonych na wykładzie otrzymuje się:

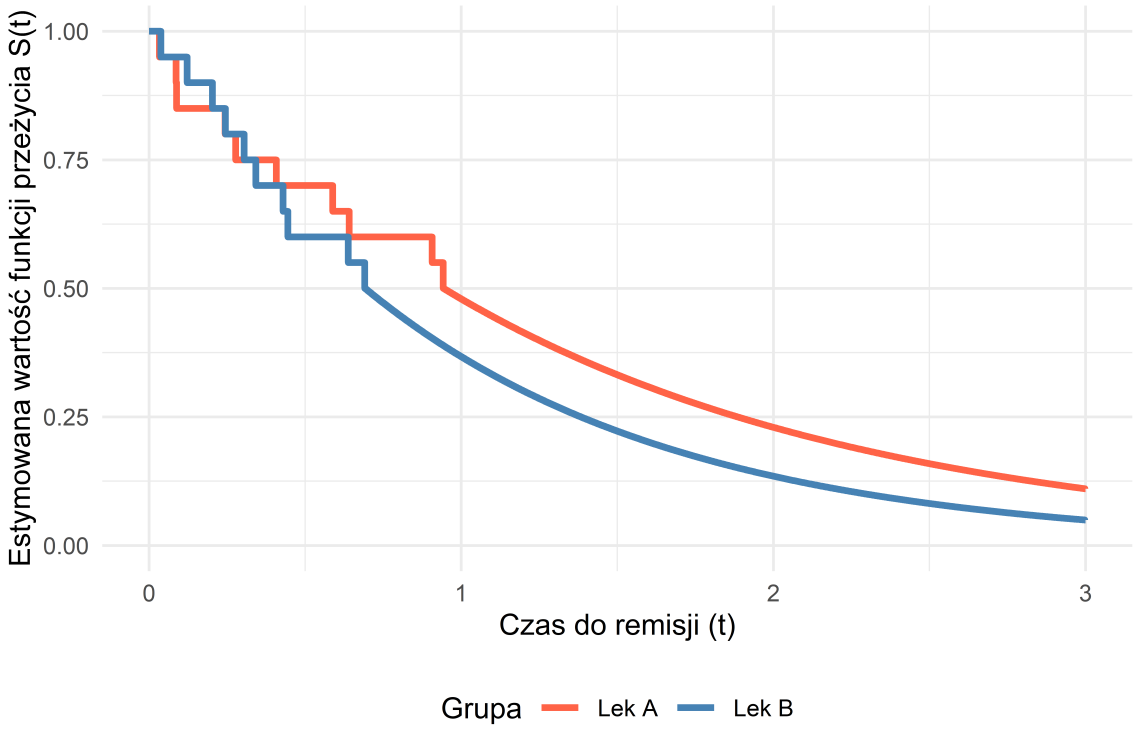
$$\hat{S}(t) = \exp\left(\frac{\ln \hat{S}(t^+)}{t^+} t\right), \quad t > t^+.$$

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)
  event_idx <- which(fit$n.event > 0)
  df.complete <- data.frame(time = fit$time[event_idx],
                            surv = fit$surv[event_idx])

  # Sprawdzenie, czy można dodać ogon (ostatni punkt jest cenzurowany)
  if (tail(fit$n.event, 1) == 0) {
    t_plus <- tail(df.complete$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.

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 licznośiach próby $n = 30, 50, 100$. Wartość t_0 przyjęto w przybliżeniu równą wartości oczekiwanej rozkładu:

$$t_0 = \lambda \cdot \Gamma\left(1 + \frac{1}{\alpha}\right),$$

gdzie $\Gamma(\cdot)$ jest funkcją 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.

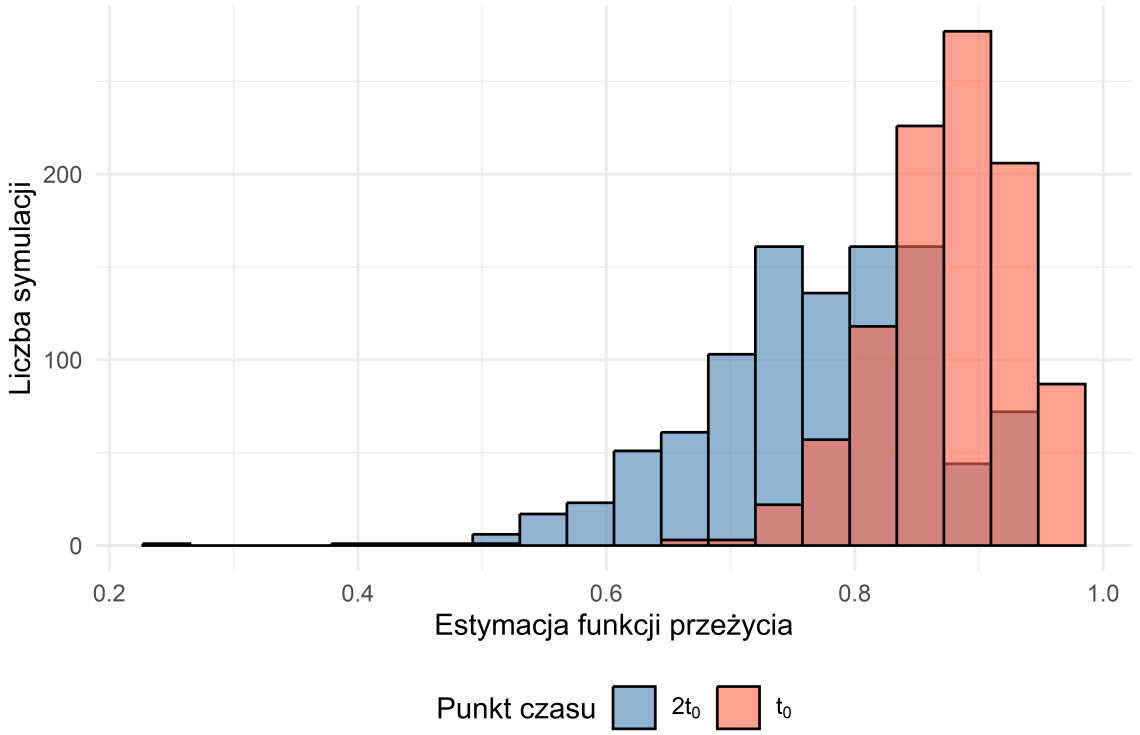
Poniżej przedstawiono blok kodu w pakiecie R wyznaczający, za pomocą estymatora Kaplana–Meiera z "ogonem", estymowaną wartość funkcji przeżycia w punkcie t.

```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])

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

  if (t <= t_plus) {
    value <- df$surv[max(which(df$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.

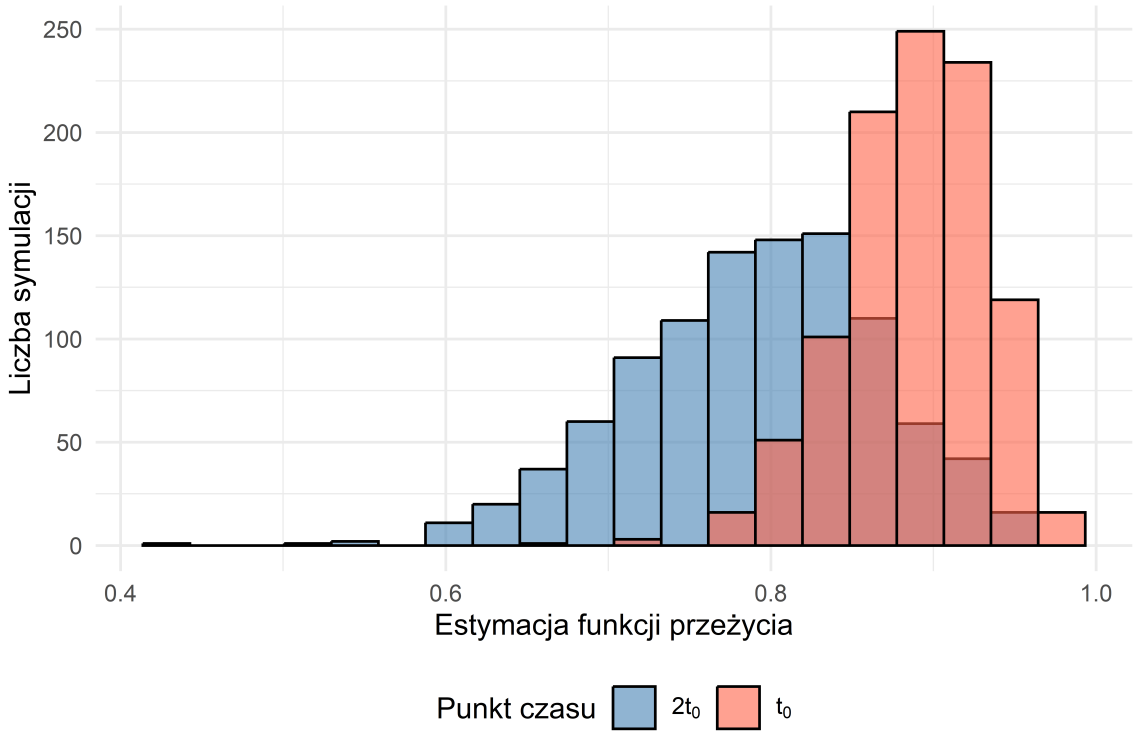
```r
M <- 10
M <- 1000 # CHUJ
n_values <- c(30, 50, 100)
alpha <- 2
lambda <- 2

t0 <- (digamma(alpha + 1) + 0.5772156649) / lambda # ja pierdole

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)

    # 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
}
```



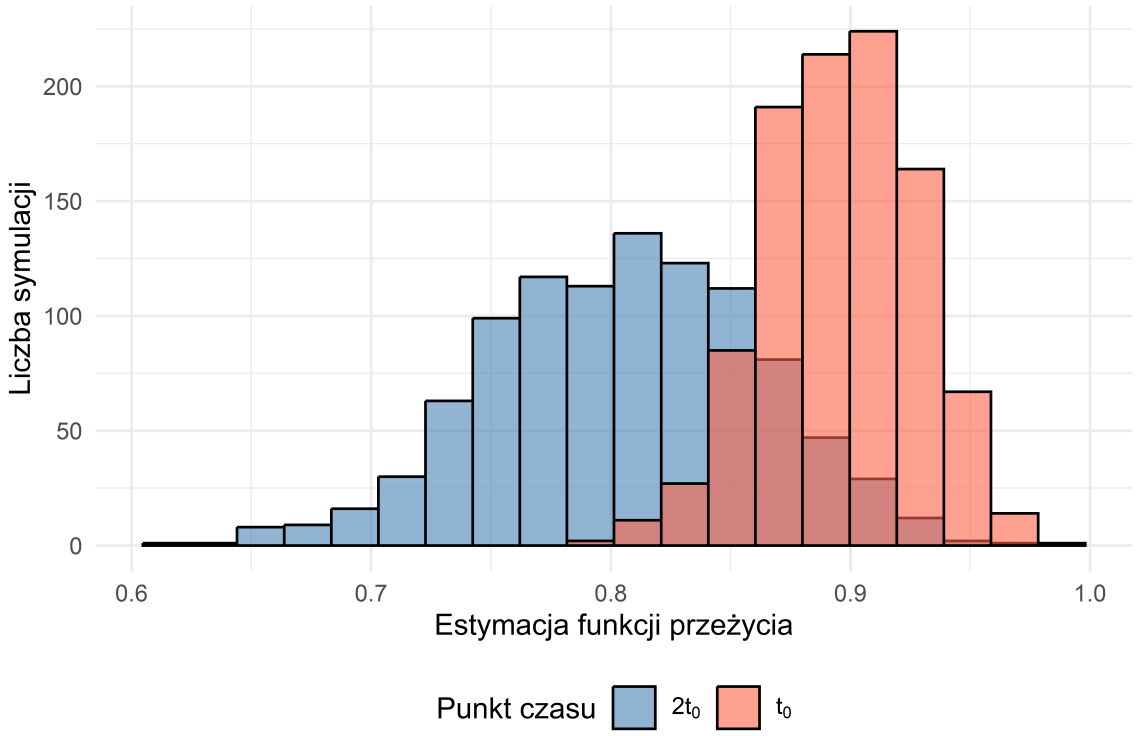
Wykres 4: Histogram oszacowań funkcji przeżycia w punktach t_0 oraz $2t_0$ - liczność próby n = 30

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.

Histogramy z poprzednich wykresów układają się w kształt dzwonowy charakterystyczny dla rozkładu normalnego, co pozwala przypuszczać, że estymator Kaplana–Meiera jest asymptotycznie normalny w punktach t_0 oraz $2t_0$.

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:

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

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

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

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

gdzie $\hat{S}(t_{(i)})$ to wartość estymatora funkcji przeżycia w chwili $t_{(i)}$, a sumowanie jest wykonywane tylko po obserwacjach do czasu ostatniej kompletnej (niecenzurowanej) 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".

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

Ostatecznie estymowana wartość oczeliwana obliczana jest ze wzoru:

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

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)

  # tylko czasy zdarzeń
  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]))

  # 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 sedniego 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 1: Średni czas życia estymowany metodami Kaplana-Meiera oraz Fleminga-Harringtona

	Lek.A	Lek.B
Kaplan-Meier	1.362	1.017
Fleming-Harrington	1.410	1.052

Tabela **??**. 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, który 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

TO DO

3.2 Zadanie 2

TO DO

4 Lista 8

Lista polega na weryfikacji hipotezy o jednakowym rozkładzie czasu do progresji choroby w dwóch badanych grupach pacjentek, przy użyciu testów log-rank, 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

TO DO

4.2 Zadanie 2

TO DO

5 Zadania dodatkowe

5.1 Zadanie 1

TO DO

5.2 Zadanie 2

TO DO

5.3 Zadanie 3

TO DO