Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka
Specjalność: –

Praca dyplomowa – licencjacka

ANALIZA PRZEŻYCIA ORAZ JEJ ZASTOSOWANIA

Monika Mrozek

słowa kluczowe:
analiza przeżycia, tabela przeżycia, wartość cenzurowana, prawdopodobieństwo przeżycia, estymator Kaplana-Meiera, logrank test, funkcja przeżycia

krótkie streszczenie:
W pracy zaimplementowano różne metody umożliwiające aproksymację prawdopodobieństwa przeżycia. Opracowano podstawowe pojęcia oraz narzędzia statystyczne wykorzystywane podczas analiz, m. in. tablice przeżycia, estymator Kaplana-Meiera oraz logrank test. Ich działanie zostało zweryfikowane dla danych wysymulowanych z różnych rozkładów prawdopodobieństwa oraz empirycznych. Opisano także przykład zastosowania tychże metod podczas badania opłacalności projektu "National Health Service Breast Screening Programme".

Opiekun pracy dyplomowej	Dr hab. inż. Marcin Magdziarz	.....	.....
	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław, rok 2019

Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Mathematics
Specialty: –

Bachelor's Thesis

SURVIVAL ANALYSIS AND ITS APPLICATIONS

Monika Mrozek

keywords:
survival analysis, life-table, censored value, probability of survival, Kaplan-Meier estimator, logrank test, survival function

short summary:
In thesis different methods were implemented which make approximation of survival probability possible. Main terms and statistic tools which are used during analysis were elaborated e.g. life-tables, Kaplan-Meier estimator and logrank test. Their working was checked for data simulated from different probability distributions and for empirical data. The example of application was also descripted during analysing the profitability of project named 'National Health Service Breast Screening Programme'.

Supervisor	Dr hab. inż. Marcin Magdziarz	.....	.....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2019

Spis treści

Spis rysunków

Spis tablic

Wstęp

Tematem pracy jest analiza przeżycia oraz jej zastosowania. Opisuje ona metody statystyczne, w których zmienną podstawowego zainteresowania jest czas trwania. Koniec tego przedziału czasowego jest jasno określony i może być wyznaczony przez różne zdarzenia – awarię urządzenia, wygaśnięcie umowy czy też zwolnienie pracownika. Czasem może być to również śmierć pacjenta – stąd pochodzenie tego terminu. Spektrum zastosowań jest więc szerokie i obejmuje wiele dziedzin, lecz tematyka tej pracy skupia się na wykorzystaniu jej metod w medycynie.

W rozdziale pierwszym przedstawiono podstawowe pojęcia oraz metody związane z analizą przeżycia. Wyjaśniono podstawowe terminy takie jak wartości cenzurowane czy zmienne prognostyczne. Następnie opisano jedną z podstawowych metod stosowanych przez statystyków, którą są tabele przeżycia. Są w nich zebrane dane na temat wielu parametrów dotyczących danej populacji takich jak prawdopodobieństwo śmierci, oczekiwana długość życia. Dla lepszego zilustrowania potrzeby korzystania z takich metod podano przykład zastosowania – dzięki tabelom przeżycia statystykom udało się odpowiedzieć na pytanie, czy opłacalne jest wykonywanie badań przesiewowych w profilaktyce raka piersi.

Funkcja przeżycia to podstawowe narzędzie, które umożliwia przewidywanie oczekiwanej długości życia pacjenta. Estymator Kaplana-Meiera pozwala na jej efektywne przybliżenie co zostało zawarte w kolejnym rozdziale. Dane poddane analizom zostały wysymulowane z różnych rozkładów za pomocą programu *RStudio*. W rozdziale tym opisano również jeden z najpopularniejszych testów statystycznych w tej dziedzinie – logrank test. Dotyczy on ryzyka śmierci w grupach pacjentów które są przedmiotem zainteresowania statystyków.

Podobne analizy zostały przeprowadzone dla danych rzeczywistych. W tym celu wykorzystano takie same narzędzia jak w przypadku danych symulowanych. Badane populacje pochodziły z trzech grup. Każda z nich była wyodrębniona na podstawie choroby, która dotknęła daną jednostkę – nowotwór złośliwy płuc, nowotwór złośliwy jelita grubego lub gammapatię monoklonalną o niezidentyfikowanym znaczeniu. Ich wybór nie był przypadkowy; w każdym z tych schorzeń prognozowany czas przeżycia jest całkowicie różny co pozwala na przyjrzenie się problemowi badawczemu z szerszej perspektywy.

Wybór takiego tematu podyktowany był moimi zainteresowaniami, którymi są różne zastosowania matematyki – szczególnie w medycynie. Myślę, że kwestie, które zostaną tu poruszone mogą być kolejnym krokiem w udowodnieniu, że matematyka to nie tylko abstrakcja jak niesłusznie sądzi wiele osób, ale także znacząca pomoc w opisywaniu zjawisk zachodzących w realnym świecie na co dzień. Medycyna i statystyka to na pozór dwie odrębne dziedziny nauki, jednak okazuje się, że istnieje bardzo wiele powiązań między nimi i rozdzielanie ich jest działaniem niepożądanym. Techniki statystyczne umożliwiają zauważenie zależności i prawidłowości w zbiorach danych, które w normalnych warunkach byłyby nie do uchwycenia. Kolejną ich zaletą jest zmniejszanie ryzyka, które ma miejsce w przypadku zastosowania nieodpowiedniego schematu leczenia. W konsekwencji można zapewnić pacjentom leczenie wyższej jakości oraz lepiej zarządzać zasobami, którymi

dysponuje służba zdrowia. Gdyby nie statystyka, medycyna rozwijałaby się zdecydowanie wolniej i postęp cywilizacyjny zostałby opóźniony. Uważam, że bez metod analizy przeżycia współczesna medycyna XXI wieku nie ma racji bytu.

Rozdział 1

Wprowadzenie

W związku z coraz szybszym rozwojem medycyny, szukamy jak najlepszych narzędzi do opisu przeprowadzonych eksperymentów klinicznych czy też badań, a także chcemy, aby ich wyniki na dużych grupach osób mogły być poddane szczegółowej analizie przede wszystkim dla rozwoju nauki, ale także dla poprawy jakości usług świadczonych przez służbę zdrowia.

Większość przeprowadzanych eksperymentów medycznych i badań ma na celu dostarczyć nowych informacji na temat prawidłowości, które można zauważyć w danej populacji. Ich zauważalną zaletą jest wielkość grupy eksperymentalnej – obserwując za każdym razem pojedynczą jednostkę ich dostrzeżenie wymagałoby zdecydowanie większej spostrzegawczości i wysiłku, a bardzo często nawet byłoby niemożliwe.

1.1 Zarys problemu i podstawowe pojęcia

Jedną z najbardziej podstawowych zmiennych zainteresowania w analizie danych medycznych bez której ciężko wyobrazić sobie jakiekolwiek badania jest długość życia pacjenta. Dokładniej określając, jest to długość przedziału czasowego, którego koniec jest wyznaczony przez określone zdarzenie – śmierć z powodu przyczyn naturalnych lub spowodowaną przez chorobę czy też nieudaną operację. Właśnie dlatego *analizą przeżycia* jest nazywana analiza tego typu danych, zebranych na przestrzeni czasu.

Analiza tego typu nie byłaby możliwa bez wykonywania eksperymentów medycznych. Może on dotyczyć leczenia nowotworu złośliwego. Załóżmy, że prognozowana długość życia jest krótka, tzn. rokowania są złe. Wspomnianym wyżej momentem końcowym będzie wtedy śmierć lub remisja choroby. Często brane pod uwagę są także jednostki, które nie osiągnęły punktu końcowego – są nadal w trakcie leczenia. Mimo, iż nie wiadomo jak długo może jeszcze żyć pacjent, pewne jest, że czas jego przeżycia musi przekroczyć ten, który upłynął do chwili obecnej; nazywamy go *wartością cenzurowaną*. Wyróżniamy trzy metody, za pomocą których można przeprowadzić procedurę *cenzurowania*, co w konsekwencji prowadzi do wyodrębnienia trzech jego rodzajów: lewego, prawego i przedziałowego [6]. Ostatnie z nich ma miejsce wtedy, kiedy wiadomo, że zdarzenie miało miejsce między datą A a B, lecz nie ma informacji na temat tego, kiedy dokładnie. W tejże pracy jest rozważane cenzurowanie prawe, ponieważ zdarzenie następuje po jakimś okresie prowadzonych badań. Lewe definuje się analogicznie do prawego.

Kolejnym ważnym pojęciem są *zmienne prognostyczne* [1]. Mamy z nimi do czynienia gdy liczba zmiennych na początku przeprowadzania eksperymentu oraz to, czy pacjent przeżyje jest silnie związane z tym, jakie przyjmą one wartości. Stosowane metody powinny uwzględniać rozkład zmiennych prognostycznych w badanych grupach oraz współpracować

z wartościami cenzurowanymi.

Analiza przeżycia jest również stosowana w badaniu śmiertelności różnych grup zawodowych. Jest to pomocne, gdy chcemy ustalić, czy narażenie na niekorzystny czynnik związany z wykonywaną pracą wpływa na większą umieralność. Na początku eksperymentu wszyscy badani są zdrowi. Następnie obserwowana jest ilość zgonów i porównywana z czasem i przyczyną jego wystąpienia, co umożliwia przeprowadzenie dokładniejszych analiz.

Bardzo wiele metod analizy opiera się na metodach korzystających z tabel przeżycia. Zostaną one opisane w kolejnym podrozdziale.

1.2 Tabele przeżycia

Miejscem, w którym zebranych jest wiele danych na temat długości życia danej grupy osób jest tabela przeżycia. Jest to jedna z najbardziej podstawowych i najstarszych metod statystycznych w dziedzinie analizy przeżycia.

Definicja 1.1 (Tabela przeżycia)**.** Tabela, która przedstawia, jakie jest prawdopodobieństwo, że osoba w danym wieku umrze przed swoimi kolejnymi urodzinami (*prawdopodobieństwo śmierci*) jest nazywana *tabelą przeżycia.*

Została wynaleziona przez angielskiego matematyka Edmonda Halley'a, który żył na przełomie XVII i XVIII wieku. Obecnie stanowi podstawę statystyki i wiedzy aktuarialnej. Aby łatwiej porównywać dane dotyczące przeżycia innych grup ludzi stosuje się metodę standaryzacji. Polega ona na grupowaniu zestawu danych ze względu na wiek badanych. Celem tworzenia takich tabel jest ukazanie ich dla danej społeczności, co pomaga w jej dokładniejszej charakterystyce.

Tabele przeżycia mogą być tworzone na różne sposoby. Najpopularniejsze i zarazem najczęściej stosowane są dwa z nich. Tabela uzyskana za pomocą jednego z nich jest nazywana bieżącą, a za pomocą drugiego – pokoleniową. Obie z tych tabel są identyczne, pod warunkiem, że populacja jest w równowadze i nie zachodzą żadne zmiany w zamieszkiwanym przez nią środowisku. W bieżącej tabeli przeżycia jest przedstawione jak kształtują się parametry związane z czasem życia przez cały jego okres. Taka tabela zawiera obecne prawdopodobieństwo zgonu dla ludzi w różnym wieku w danym roku. Metody pozwalające je tworzyć są dosyć skomplikowane, lecz stosowane są tylko w biurach ubezpieczeniowych lub do opracowania ważnych danych państwowych. Dla ułatwienia inne, bardziej trywialne rozwiązanie zostało zaproponowane niecałe 10 lat później, w roku 1991.

Aby stworzyć tabelę pokoleniową, również należy zwracać uwagę na pewne obostrzenia. Badani, dla których będzie ona utworzona, muszą być urodzeni we wcześniej określonym przedziale czasowym. Przedstawia się w niej wskaźniki śmiertelności w całym okresie życia pewnej populacji.

Przykładowe dane zostały ukazane w tabeli 1.1. Prawdopodobieństwo, że badany w danym wieku (oznaczonym przez x) umrze przed swoimi kolejnymi urodzinami można znaleźć w drugiej kolumnie. Zmienna zarezerwowana dla tej wartości to q_x. Liczbę pacjentów, którzy nadal żyją oznaczamy przez l_x – jest ona wybrana z 1000 losowo wybranych urodzeń żywych. Wyrażając to w sposób bardziej precyzyjny [1]

$$l_x = l_0 \; p_0 \; p_1 \ldots p_{x-1} \tag{1.1}$$

Podane wartości należy oczywiście odczytać z tabeli 1.1. Jedna z nich nie jest podana bezpośrednio, lecz uzyskano ją z prostego wzoru $p_x = 1 - q_x$. Sprawdzimy to dla

najprostszego przypadku, gdy x=1. Dla tej sytuacji $l_1 = p_0\ l_0$. Przeprowadzając dokładne obliczenia otrzymujemy:

$$l_1 = p_0\ l_0 = (1 - q_0)\ l_0 = (1 - 0.0719)\ 1000 = 928.1$$

Zgadza się to z wartością, którą można znaleźć w tabeli.

Tabela 1.1: Bieżąca i pokoleniowa tabela przeżycia dla mężczyzn w Anglii i Walii urodzonych w roku 1931 (wersja skrócona)

	Bieżące tabele przeżycia 1930-32			Pokoleniowa tabela przeżycia, 1931
Wiek (w latach) x	Prawodpodobieństwo śmierci pomiędzy wiekiem x a $x+1$ q_x	Pacjenci, którzy przeżyli l_x	Oczekiwana długość życia e_x	Pacjenci, którzy przeżyli l_x
0	0.0719	1000	58.7	1000
1	0.0153	928.1	62.2	927.8
5	0.0034	900.7	60.1	903.6
10	0.0015	890.2	55.8	894.8
20	0.0032	872.4	46.8	884.2
30	0.0034	844.2	38.2	874.1
40	0.0056	809.4	29.6	861.8
50	0.0113	747.9	21.6	829.7
60	0.0242	636.2	14.4	–
70	0.0604	433.6	8.6	–
80	0.1450	162.0	4.7	–

Źródło: Statistical methods in medical research, P. Armitage, G. Berry, J.N.S. Matthews [1]

Czwarta kolumna przedstawia wartość e_x, która jest średnią długości pewnych przedziałów, które są wyznaczone przez czas życia od momentu osiągnięcia wieku x. Jest ona liczona dla wszystkich pacjentów l_x, którzy dożyli tego wieku. Patrząc na problem globalnie, tzn. dla całej tabeli, w przybliżeniu można wyliczyć tę wartość ze wzoru:

$$e_x = (l_{x+1} + l_{x+2} + \ldots)/l_x + \frac{1}{2} \tag{1.2}$$

W nawiasie zliczamy liczbę lat przeżytych przez pacjentów l_x w powyżej określonym przedziale czasowym. Ich zgon nastąpił natychmiast po y-wych urodzinach, przy założeniu, że miał miejsce pomiędzy wiekiem y a $y+1$. Ułamek dopisany na końcu pełni rolę korekty wyniku końcowego, z powodu faktu, iż zgon może nastapić w każdym momencie roku. Niestety jest to dość niedokładne oszacowanie.

Jak wcześniej wspomniano, pokoleniowe tabele przeżycia dotyczą osób, które urodziły się w tym samym czasie – stąd poniekąd wywodzi się jej nazwa. Można powiedzieć, że analizują dane przyszłościowo, ponieważ dostarczają danych na temat śmiertelności

ludzi w różnym wieku, lecz dopiero dla momentu, kiedy zostałby on osiągnięty. Wartości ukazano w ostatniej kolumnie tabeli 1.1. Rok urodzin badanych obejmuje okres pięciu lat i oscyluje wokół roku 1931.

Jak można wywnioskować, wartość l_1 powinna być zbliżona dla obu rodzajów tabel, ponieważ dane pochodzą z tego samego okresu. Jest wyliczona na podstawie śmiertelności noworodków. Wraz ze wzrostem x, zauważa się, że l_x jest większy dla tabeli pokoleniowej tzn. w późniejszym wieku jest większa dysproporcja co do liczby badanych, którzy przeżyli. Jest to spowodowane opieraniem się o wskaźniki śmiertelności dla późniejszych okresów, a jak wiadomo wraz z upływem czasu i postępem cywilizacyjnym, średnia długość życia się wydłuża. Tabela potwierdza również oczywiste fakty – im badany jest starszy, tym większe prawdopodobieństwo jego śmierci, mniejsza ilość osób, które przeżyły, krótsza oczekiwana długość życia. Prawdopodobieństwo śmierci dla osiemdziesięciolatka jest już dość duże i wynosi 0.1450. Dla porównania w wieku niemowlęcym wynosi 0.0719.

Tabele przeżycia używane są nie tylko w demografii i statystyce, ale także w wielu innych dziedzinach. Często są tworzone dla oszacowania opłacalności różnych projektów. Jednym z nich był *National Health Service Breast Screening Programme (NHSBSP)* [2]. Miał na celu pomoc w ocenie, czy wykonywanie badań screeningowych ma sens. Naukowcy poddali badaniom trzy hipoteteczne grupy kobiet. Wszystkie z nich miały 50 lat i u żadnej z nich nie stwierdzono przypadku zachorowania przez ostatnie 35 lat. W pierwszej grupie nie były przeprowadzane żadne badania przesiewowe. Badane z drugiej z nich miały wykonywaną regularnie mammografię co trzy lata począwszy od pięćdziesiątego roku życia, kończąc na sześćdziesiątym dziewiątym. Ostatnia grupa została poddana takiemu samemu schematowi jak druga, lecz z pewną różnicą; otrzymały je tylko te panie, u których stwierdzono za pomocą metody estymacji większe ryzyko zachorowania. Wszystkie przeprowadzane analizy i badania miały miejsce w okresie od lipca 2016r. do września 2017r.

Do przeprowadzenia badania niezbędne było przyjęcie odpowiedniego modelu rozkładu prawdopodobieństwa dla ryzyka zachorowania. Przyjęto wariancję na poziomie 0.43. Została ona wyliczona na podstawie:

- średniej wartości wariancji; średnia została wyliczona z przypadków, gdzie odpowiednio statystycy mieli informację lub nie na temat umiejscowienia zmutowanego genu nowotworowego, który odpowiadał za pojawienie się choroby,
- łącznej wariancji, na którą miały wpływ czynniki epidemiologiczne ryzyka, czyli ogół postępowań, które mogą prowadzić do rozwoju danej choroby, oraz zmutowane geny, których umiejscowienie w genomie było znane.

Przyjęto, że interakcja między czynnikami genetycznymi a epidemiologicznymi ma model log-addytywny. W konsekswencji rozkład prawdopodobieństwa dla ryzyka zachorowania jest lognormalny. Poniżej można znaleźć wzór na jego gęstość.

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

Mając daną średnią i wariancję dla takiego rozkładu można obliczyć wiele parametrów opisujących ten model. Przykładowo dla danego ryzyka można policzyć pozycję percentylową. Jest to procent wyników, który jest powyżej lub poniżej niego patrząc na rozkład częstości. Liczony jest ze wzoru:

$$\frac{c_\ell + 0.5 f_i}{N} \cdot 100\% \tag{1.3}$$

gdzie:
c_ℓ - liczba wyników, które są mniejsze od ustalonego progu
f_i - częstość dla ustalonego progu (wartości)
N - ilość badanych w próbie

Oszacowano jaka część populacji jest posiadaczami genu, który odpowiada za większe ryzyko zachorowania niż ustalony próg oraz ile może pojawić się przypadków ekspresji genu odpowiadającego za rozwój choroby w obrębie grupy podwyższonego ryzyka.

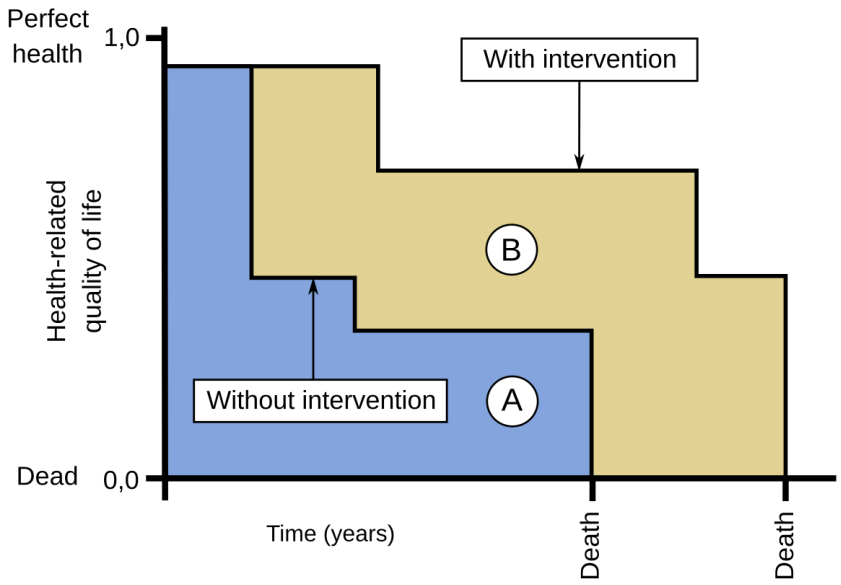
Dla utworzonych wcześniej grup przeprowadzono dokładną analizę kosztów. Użyto w niej wskaźnika ICER (*incremental cost-effectiveness ratio*). Przedstawia on różnicę w średnich kosztach, które generowały grupy. W jednej z nich panie wykonywały regularnie badania przesiewowe, zaś w drugiej nie korzystały z tego typu świadczeń medycznych. Podział został wykonany na podstawie średniej QALY (*quality-adjusted life-years*). Jest to jeden ze wskaźników powszechnie stosowanych w analizie przeżycia.

Definicja 1.2 (QALY)**.** *QALY* to wskaźnik liczby lat życia ściśle powiązanych z jego jakością. Pomaga analizować skuteczność procedur medycznych stosowanych w ochronie zdrowia. Wynik jest uzyskiwany poprzez przemnożenie dwóch wartości:

- liczby lat, którą przeżyje pacjent dzięki zastosowanemu schematowi leczenia (przykładowo w przypadku 9 lat, będzie wynosić 9)
- liczby z przedziału [0,1], która określa zadowolenie pacjenta z uzyskanej jakości życia w wyniku leczenia; rośnie proporcjonalnie do satysfakcji. Bierze pod uwagę zarówno stan zdrowia, jak i samopoczucie psychiczne.

Wykres jakości życia od uzyskanego w konsekwencji czasu życia przez pacjenta przedstawiono na rysunku 1.1.

Źródło: `https://en.wikipedia.org`



Rysunek 1.1: QALY dla grup pacjentów poddanych leczeniu oraz tych, którzy nie otrzymali żadnej pomocy medycznej

W celu uzyskania jeszcze większej ilości informacji na temat badanych populacji, zostały dla nich wyliczone wskaźniki QALY.

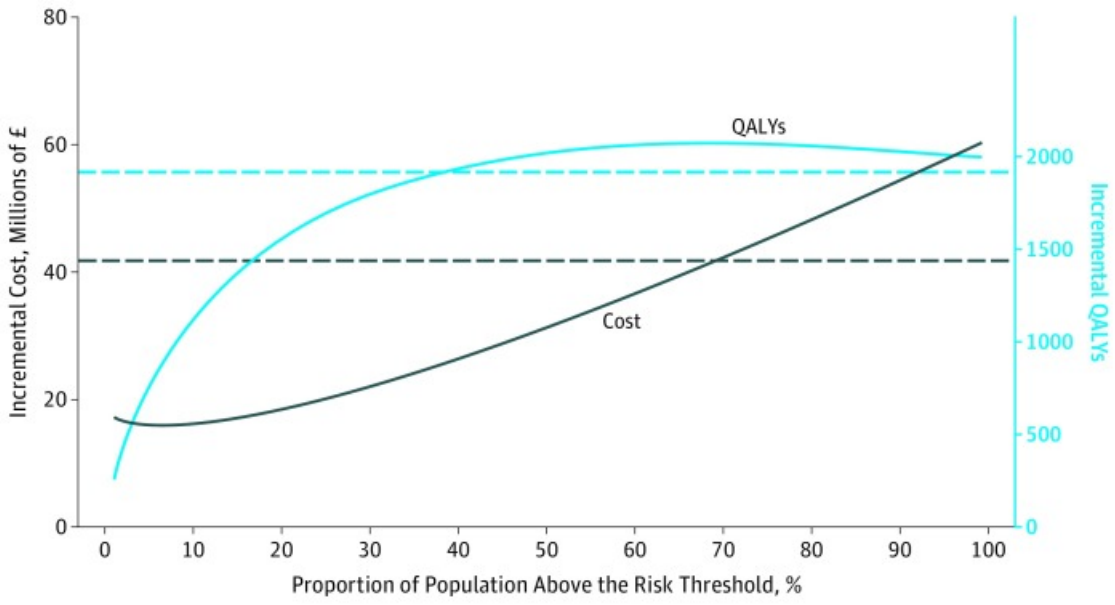
Trzecia z badanych grup zawierała 99 schematów postępowania w zależności od ryzyka zachorowania; średnia wartość *quality-adjusted life-years* została pomnożona przez odpowiadający mu WTP (*willingness to pay*, w ekonomii jest to maksymalna kwota jaką jest w stanie zapłacić konument za określone dobro lub usługę), który został pomniejszony o koszt całkowity.

Ponadto dla każdej grupy kobiet obliczono NMB (*Net Monetary Benefit*). Schemat leczenia z najwyższą jego wartością dla danego WTP był uznawany za najbardziej korzystny, ponieważ za kryterium wyboru uznano wymagane nakłady finansowe.

Rezultaty przeprowadzonych badań można odczytać z wykresu 1.2. Głównym obszarem zainteresowania naukowców były grupa pierwsza i trzecia, dlatego też to one zostały porównane. Odnoszą się do nich wykresy przedstawione za pomocą linii ciągłych. Lewa oś obrazuje narastający koszt w milionach funtów, a prawa – wzrost wartości wskaźników QALY. Liniami przerywanymi zostały zaznaczone wzrost wysokości kosztów oraz wartości QALY dla drugiej grupy (poddawanej badaniom screeningowym w wieku 50-69 lat); porównane zostały z grupą niewykonującą takich badań. Część populacji, która znajduje się powyżej progu ryzyka obliczamy pomniejszając liczbę sto o ryzyko wyrażone w percentylach dla jego rozkładu prawdopodobieństwa (oś pozioma). Zgodnie z tą regułą przy zerowym nakładzie kosztów cała populacja jest zagrożona, lecz w momencie gdy będą one wynosić około 60 milionów funtów, można założyć, że każdy pacjent przeżyje.

Oczywiście, aby stwierdzić, która ze strategii jest najbardziej efektywna wykonano jeszcze szereg innych analiz. Ostatecznie okazało się, że najbardziej właściwa jest ta, która zakłada wykonywanie badań przesiewowych w zależności od ryzyka zachorowania, a więc ta, której była poddana grupa trzecia. To wszystko nie byłoby możliwe do osiągnięcia, gdyby nie metody opierające się na danych które są zebrane w tabelach przeżycia.

Źródło: `https://europepmc.org/articles/PMC6230256`



Rysunek 1.2: Wzrost kosztów i wartości QALY dla grupy poddawanej badaniom w zależności od ryzyka zachorowania w porównaniu do grupy, która ich nie wykonywała

Rozdział 2

Analiza przeżycia - modele, ich własności i estymatory

2.1 Estymator Kaplana-Meiera

Estymator to jeden z najczęstszych terminów używanych w statystyce. Pomaga on dowiedzieć się, jaka jest wartość parametru rozkładu np. w wybranej populacji. Może być nim przykładowo średnia z próby. Tutaj zostanie opisany estymator stosowany powszechnie w analizie przeżycia – estymator Kaplana-Meiera. Jego nazwa pochodzi od nazwisk dwóch naukowców, którzy w 1958 roku opracowali tę metodę estymacji. W omawianym tu przypadku parametrem zainteresowania będzie frakcja pacjentów, którzy w dalszym ciągu żyją. Estymowana zaś zostanie funkcja przeżycia. W literaturze angielskiej można spotkać się również z nazwą *product-limit estimate* [4].

W początkowej fazie eksperymentu dane dzieli się na jak najmniejsze przedziały czasowe – tak małe, jak pozwala na to ich dokładność. Dzięki temu uzyskane rezultaty będą bardziej wiarygodne, niż dla większych odstępów czasowych. Jeżeli obliczenia mają zostać przeprowadzone z precyzją co do jednego dnia, należy zebrać dane, które nie są mniej dokładne od zażądanej. Długości przedziałów czasowych są przez nie wyznaczone; innymi słowy precyzja co do dnia determinuje jednodniowe przedziały, a co do dwóch dni – dwudniowe. Jeśli przyjmie się, że w momencie czasowym t_j miało miejsce d_j zgonów oraz, że na krótko przed tym zdarzeniem żyło n'_j pacjentów to przybliżone prawdopodobieństwo śmierci dokładnie w tym czasie oblicza się ze wzoru [1]:

$$q_{t_j} = d_j / n'_j \tag{2.1}$$

Dla większości przedziałów d_j=0, co implikuje q_{t_j}=0. Prawdopodobieństwo przeżycia $p_{t_j} = 1 - q_{t_j}$ wynosi wtedy 1. Istnieje jednak możliwość ich pominięcia dzięki zastosowaniu wzoru $l_x = l_0\ p_0\ p_1 \ldots p_{x-1}$. W konsekwensji przedstawionych tutaj rozważań estymatorem przeżycia w chwili t jest [1]:

$$l_t = \prod_j p_{t_j} = \prod_j \frac{n'_j - d_j}{n'_j} \tag{2.2}$$

W produkcie uwzględniane są wszystkie przedziały czasowe, w których nastąpił zgon. Postępowanie trwa aż do chwili mniejszej bądź równej t. Co więcej, estymator ten jest estymatorem MLE (*maximum likelihood estimator*).

Definicja 2.1 (Estymator największej wiarogodności)**.** Załóżmy, że $X_1, X_2, \ldots, X_n$ i.i.d. pochodzą z populacji o gęstości (funkcji prawdopodobieństwa) f(x, θ), gdzie $\theta \in \Theta$ jest nieznanym parametrem. Określamy funkcję wiarogodności

$$L(\theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta).$$

Estymator największej wiarogodności parametru θ to taka $\hat{\theta} \in \Theta$, dla której $L(\theta)$ osiąga największą wartość, tzn:

$$\hat{\theta} = \underset{\hat{\theta} \in \Theta}{\operatorname{argmax}} L(\theta)$$

Działanie estymatora Kaplana-Meiera najlepiej ilustrują wykresy. Można zastosować go dla danych symulowanych, a także medycznych.

2.2 Estymator Kaplana-Meiera dla danych symulowanych

Symulacja danych będzie się odbywać w oparciu o trzy popularne rozkłady – wykładniczy, Weibulla oraz Pareto. Badania zostaną przeprowadzone dla 1000 osób, co oznacza, że taki będzie rozmiar rozważanej próby. Do każdego z pacjentów jest przypisany czas przeżycia od jakiegoś charakterystycznego dla niego momentu np. operacji czy zdiagnozowania danej choroby. Oczywiście tak jak wcześniej, stosowany jest proces cenzurowania. Znajduje on zastosowanie, gdy czas badania kończy się przed wystąpieniem określonego zdarzenia, ponieważ w takiej sytuacji nie ma informacji na temat długości przedziału czasowego od zakończenia badania do danego zdarzenia. Oprócz czasu przeżycia każdego badanego, jest dostępna również bazowa informacja – czy nadal żyje. Inaczej patrząc, przez $P(T > t)$ można oznaczyć prawdopodobieństwo tego, że pacjent będzie żył w dalszym ciągu po chwili t, gdzie T jest zmienną losową mierzącą czas życia.

W pierwszym przypadku czas przeżycia w latach będzie pochodził z rozkładu wykładniczego z parametrem $\lambda = 1/10$ tzn. $exp(1/10)$. Gęstość tego rozkładu jest funkcją ciągłą i jest dana wzorem:

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geqslant 0 \end{cases} \tag{2.3}$$

Wartość średnią ET uzyskuje się za pomocą prostego rachunku – obliczenia całki niewłaściwej:

$$ET = \int_{-\infty}^{\infty} t f(t) dt = \int_0^{\infty} t\lambda e^{-\lambda t} dt = \left[-e^{-\lambda t}(t + \frac{1}{\lambda}) \right]_0^{\infty} = 0 - \left(-\frac{1}{\lambda} \right) = \frac{1}{\lambda} \overset{\lambda = 1/10}{=} 10.$$

Wykres estymatora został zaznaczony kolorem czerwonym. W celu sprawdzenia, w jaki działa on sposób, zielonym kolorem przedstawiono wykres $P(T > t)$, czyli funkcji przeżycia.

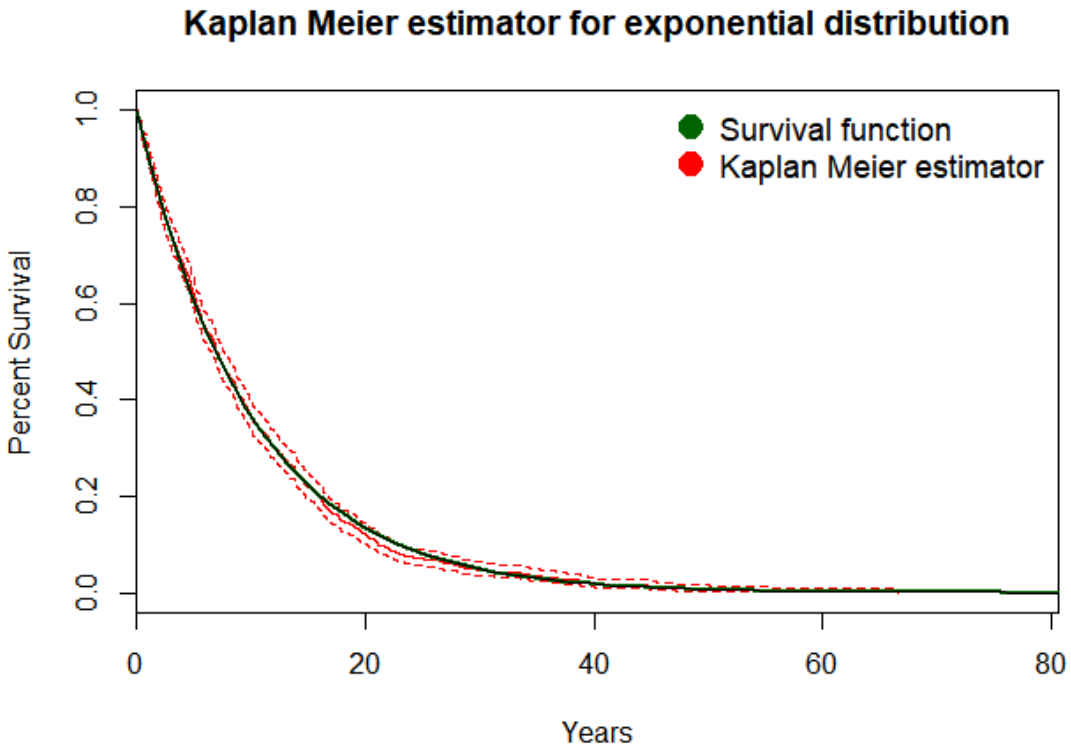
Łatwo widać, że w przypadku rozkładu wykładniczego

$$P(T > t) = 1 - P(T < t) = 1 - F(t) = e^{-\lambda t} \tag{2.4}$$

gdzie $F(t)$ jest dystrybuantą.

Rezultat analizy dla tysiąca obserwacji i opisanego wyżej rozkładu przedstawiono na rysunku 2.1. Dzięki rachunkom (2.4) widać związek między funkcją przeżycia oraz dystrybuantą. Im więcej obserwacji, tym większe dopasowanie pomiędzy obydwoma

Źródło: Opracowanie własne



Rysunek 2.1: Estymator Kaplana-Meiera oraz funkcja przeżycia dla rozkładu wykładniczego

wykresami na omawianym rysunku, ponieważ estymator Kaplana-Meiera w granicy dąży do funkcji przeżycia. Przedział ufności wynosi 95% i został zaznaczony na wykresie liniami przerywanymi. Analogiczne oznaczenia zostały zastosowane dla kolejnych wykresów.

Kolejnym rozkładem prawdopodobieństwa, z którego zostały wysymulowane dane jest rozkład Weibulla z parametrem kształtu $\alpha > 0$ i skali $\beta > 0$ o gęstości:

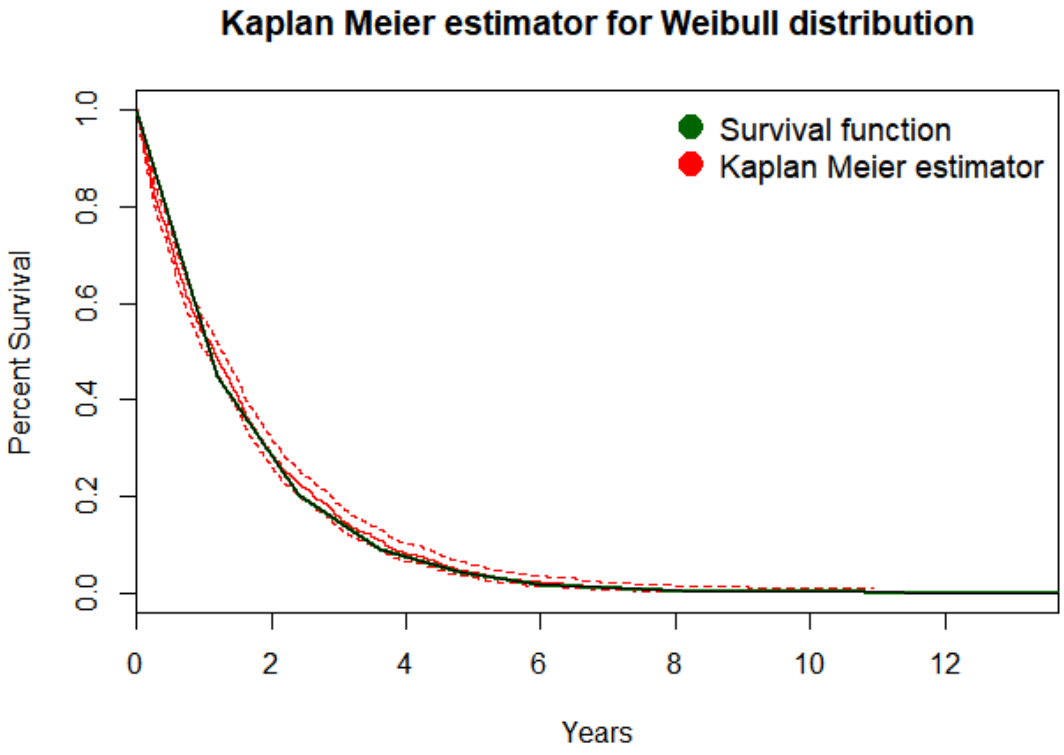
$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] & x \geqslant 0 \end{cases} \tag{2.5}$$

Jest to jeden z możliwych wzorów, ponieważ istnieje kilka ich wariantów. W innych źródłach można napotkać na następującą wersję:

$$f(x) = \begin{cases} 0 & x < 0 \\ \alpha\beta x^{\beta-1} e^{-\alpha t^{\beta}} & x \geqslant 0 \end{cases} \tag{2.6}$$

Analogiczna sytuacja ma miejsce przykładowo dla rozkładu $\Gamma(\alpha, \beta)$, gdzie również jest do wyboru kilka wersji wzoru zapisu gęstości. Tutaj jednak zostanie wykorzystany pierwszy z nich, ponieważ jest częściej stosowany. Kluczową obserwacją jest, że w przypadku rozkładu wykładniczego lata przeżycia były zdecydowanie dłuższe; tutaj dla $\alpha = 1$ i $\beta = 3/2$ prawdopodobieństwo przeżycia zbliża się do zera już w okolicy 8 lat od diagnozy, co można odczytać z wykresu 2.2. Taki rozkład może być wykorzystywany w modelowaniu przeżycia dla pacjentów chorujacych na niezbyt dobrze rokujące schorzenia np. nowotwór złośliwy płuc.

Źródło: Opracowanie własne



Rysunek 2.2: Estymator Kaplana-Meiera oraz funkcja przeżycia dla rozkładu Weibulla

Warto sprawdzić, jaką wartość oczekiwaną ma taki rozkład:

$$ET = \int_{-\infty}^{\infty} tf(t)dt = \int_0^{\infty} t\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} exp\left[-\left(\frac{t}{\beta}\right)^{\alpha}\right]dt = \left|\begin{array}{c}\left(\frac{t}{\beta}\right)^{\alpha}=u \\ t=\beta u^{1/\alpha} \\ \frac{\alpha t^{\alpha-1}}{\beta^{\alpha}}dt=du\end{array}\right| =$$

$$= \int_0^{\infty} \beta u^{1/\alpha} e^{-u}du = \beta\int_0^{\infty} u^{1/\alpha+1-1}e^{-u}du \overset{(*)}{=} \beta\Gamma(1+\frac{1}{\alpha}) \overset{(**)}{=} \beta\frac{1}{\alpha}\Gamma(\frac{1}{\alpha}) \overset{\alpha=1}{\underset{\beta=3/2}{=}} \frac{3}{2}\cdot 1 = \frac{3}{2}$$

W przejściu oznaczonym $(*)$ zastosowano wzór $\Gamma(a) = \int_0^{\infty} u^{a-1}e^{-u}du$, a $(**)$ ważną własność funkcji gamma – $\Gamma(a+1) = a\Gamma(a)$. Wartość oczekiwana wynosząca $3/2$ zgadza się z tym co można odczytać z wykresu 2.2.

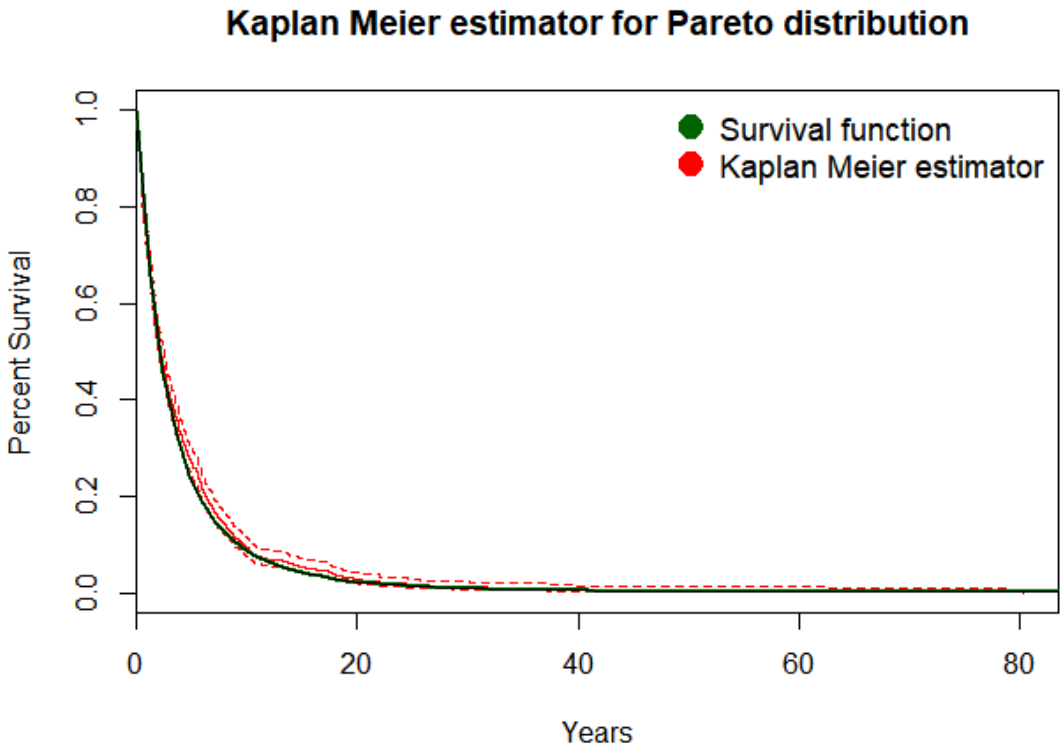
Trzecim rozkładem prawdopodobieństwa, dla którego przedstawiono funkcję przeżycia oraz estymator Kaplana-Meiera jest rozkład Pareto $Pa(k, \alpha)$, przy czym zakładamy, że $\alpha > 1$. Efekty analiz zostały ukazane na rysunku 2.3. Wybrane parametry rozkładu to $k = 4, \alpha = 3$.
W tym przypadku gęstość dla $x > k$ wynosi

$$f(x) = \frac{\alpha k^{\alpha}}{x^{\alpha+1}} \tag{2.7}$$

Zatem:

$$ET = \int_{-\infty}^{\infty} tf(t)dt = \int_k^{\infty} t\frac{\alpha k^{\alpha}}{t^{\alpha+1}}dt = \int_k^{\infty} \frac{\alpha k^{\alpha}}{t^{\alpha}}dt = \alpha k^{\alpha}\int_k^{\infty}\frac{1}{t^{\alpha}}dt \overset{(*)}{=} \alpha k^{\alpha}\left[\frac{t^{-\alpha+1}}{-\alpha+1}\right]_k^{\infty} =$$

Źródło: Opracowanie własne



Rysunek 2.3: Estymator Kaplana-Meiera oraz funkcja przeżycia dla rozkładu Pareto

$$= \alpha k^{\alpha} \left(-\frac{k^{-\alpha+1}}{-\alpha+1} \right) = \frac{\alpha k}{\alpha - 1} \overset{\substack{k=4 \\ \alpha=3}}{=} \frac{3 \cdot 4}{2} = 6$$

W przejściu oznaczonym $(*)$ konieczne było założenie, że $\alpha > 1$ – w przeciwnym wypadku całka byłaby rozbieżna i dalsze obliczenia traciłyby sens. Otrzymano więc, że oczekiwana wartość życia wynosi 6 lat, czyli mniej niż w przypadku rozkładu wykładniczego, ale więcej niż dla Weibulla.

2.3 Logrank test

Początki historii testowania hipotez sięgają lat trzydziestych ubiegłego wieku. Matematyczna teoria dotycząca tej procedury została wynaleziona przez dwóch znanych statystyków – Jerzego Neymana i Karla Pearsona [3]. Wśród wielu testów statystycznych, logrank test jest tym, który odgrywa ważną rolę w analizie przeżycia, jednakże nie jest jedynym; alternatywą dla niego jest Peto test [5]. Logrank test znajduje zastosowanie nie tylko dla dwóch grup danych, ale także dla ich większej ilości. Są one zestawiane w tabelach – osobno dla każdej z nich. Jej układ można zobaczyć poniżej.

W tabeli 2.1 przedstawione zostały dane na temat dwóch grup. Punkt odniesienia to chwila t_j – liczba zgonów pochodzi właśnie z tego okresu, a ilość pacjentów, którzy przeżyli odnosi się do chwili mającej miejsce tuż przed t_j.

Przypuśćmy, że hipoteza zerowa brzmi *ryzyko śmierci jest takie samo w obu grupach*, przy czym hipoteza alternatywna jest zaprzeczeniem zerowej, tzn. dla obu populacji

Tabela 2.1: Tabela z danymi (logrank test)

	Liczba zgonów	Liczba pacjentów, którzy przeżyli	Razem
Grupa A	d_{jA}	$n'_{jA} - d_{jA}$	n'_{jA}
Grupa B	d_{jB}	$n'_{jB} - d_{jB}$	n'_{jB}
Razem	d_j	$n'_j - d_j$	n'_j

Źródło: Statistical methods in medical research, P. Armitage, G. Berry, J.N.S. Matthews [1]

ryzyko jest różne. Przy spełnieniu założeń pierwszej z nich rozkład liczby zgonów powinien spełniać w każdym momencie czasowym poniższe warunki [1]

$$\begin{cases} E(d_{jA}) = n'_{jA} d_j / n'_j \\ var(d_{jA}) = \dfrac{d_j (n'_j - d_j) n'_{jA} n'_{jB}}{{n'_j}^2 (n'_j - 1)} \end{cases} \tag{2.8}$$

Przyjmując d_j=1, równania 2.8 przyjmują prostszą postać:

$$\begin{cases} E(d_{jA}) = n'_{jA}/n'_j = p'_{jA} \\ var(d_A) = p'_{jA}(1 - p'_{jA}) \end{cases}$$

Jak zapisano powyżej, stosunek pacjentów, którzy przeżyli w grupie A oznacza się p'_{jA}. Wartość d_{jA} oraz $E(d_{jA})$ są różne, co jest dowodem przeciwko hipotezie zerowej. Kombinację takich różnic, które pojawiały się w różnym czasie nazywamy właśnie logrank test. Jego pierwowzorem był test Mantela-Haenszela. Powstał w 1966 roku i można odnaleźć w nim wiele analogii. Z powodu mnogości przedziałów (sytuacja dyskretna) zachodzą wzory [1]

$$\begin{aligned} O_A &= \sum d_{jA} \\ E_A &= \sum E(d_{jA}) \\ V_A &= \sum var(d_{jA}) \end{aligned} \tag{2.9}$$

Powyższe sumowania wykonuje się po t_j. Ponadto zachodzi równość E_A+E_B = O_A+O_B, co można sprawdzić prostym rachunkiem:

$$\begin{aligned} E_A + E_B &= O_A + O_B \\ \sum E(d_{jA}) + \sum E(d_{jB}) &= \sum d_{jA} + \sum d_{jB} \\ \sum \frac{n'_{jA} d_j}{n'_j} + \sum \frac{n'_{jB} d_j}{n'_j} &= \sum d_{jA} + \sum d_{jB} \\ \sum n'_{jA} d_j + \sum n'_{jB} d_j &= \sum d_{jA} n'_j + \sum d_{jB} n'_j \\ \sum d_j (n'_{jA} + n'_{jB}) &= \sum n'_j (d_{jA} + d_{jB}) \end{aligned}$$

Równanie to jest tożsamościowe, ponieważ $n'_{jA} + n'_{jB} = n'_j$ oraz $d_{jA} + d_{jB} = d_j$. Statystyka testowa stosowana do weryfikacji sformułowanej wcześniej hipotezy to [1]

$$\chi^2_1 = \frac{(O_A - E_A)^2}{V_A} \tag{2.10}$$

co jest dokładnie rozkładem chi-kwadrat z jednym stopniem swobody ($\chi^2_{(1)}$). Aby uniknąć liczenia wariancji, można skorzystać z równoważnego wzoru:

$$\chi^2_2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \tag{2.11}$$

Tę statystykę testową można wykorzystać również dla przypadku, gdy liczba grup jest większa od dwóch. W ogólnej sytuacji, gdy mamy k grup, test statystyczny ma $k - 1$ stopni swobody; gdy $k = 2$ jest to jeden, co zgadza się ze wcześniejszymi rozważaniami. Z opisywanym tutaj testem statystycznym nierozłącznie związany jest stosunek ryzyka (j. ang. *hazard ratio*), Oznaczany jest literą h i wyraża się wzorem [1]

$$h = \frac{O_A/E_A}{O_B/E_B} \tag{2.12}$$

Powyższy iloraz estymuje wskaźnik śmiertelności w grupie A (porównując do grupy B). Ponadto zachodzi

$$SE[ln(h)] = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}}$$

gdzie SE to błąd standardowy. Kiedy stosunek ryzyka (2.12) nie jest bliski jedności, wtedy przedziały ufności dla błędów standardowych będą miały bardzo mały zakres. Ponadto jest on obciążony.

Logrank test jest testem nieparametrycznym. Testy tego typu są wykorzystywane gdy nie są spełnione założenia o normalności czy też jednorodności (inaczej homogeniczności) wariancji. Drugie z nich oznacza równość wariancji; bardziej obrazowo – różnorodność danych w każdej z grup jest podobna. Konsekwencją tego jest ich mniejsza moc w porównaniu do mocy testów parametrycznych.

Definicja 2.2 (Moc testu)**.** Prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona fałszywa nazywamy *mocą testu* statystycznego. Innymi słowy, jest to prawdopodobieństwo niepopełnienia błędu drugiego rodzaju. [7]

W pozyższej definicji odwołano się do pojęcia błędu występującego podczas testowania hipotez. Jest to jeden z dwóch ich rodzajów.

Definicja 2.3 (Błąd pierwszego rodzaju)**.** *Błąd pierwszego rodzaju* występuje, gdy odrzucana jest hipoteza zerowa mimo, że jest prawdziwa, przy czym przyjmowana jest hipoteza alternatywna. [9]

Definicja 2.4 (Błąd drugiego rodzaju)**.** *Błąd drugiego rodzaju* występuje, gdy przyjmowana jest hipoteza zerowa mimo, że hipoteza alternatywna jest prawdziwa, przy czym jest ona odrzucana. [9]

Zatem moc testu to prawdopodobieństwo tego, że nie zostanie przyjęta hipoteza zerowa, gdy prawdziwa jest hipoteza alternatywna.

2.4 Logrank test dla danych symulowanych

Przypomnijmy, że głównym problemem, który pomaga rozwiązać logrank test jest rozstrzygnięcie, czy krzywe przeżycia dla analizowanych grup pokrywają się. Wynika to z hipotezy zerowej, która jest weryfikowana w tym teście. Zakłada ona, jak wyjaśniono we wcześniejszej części, że prawdopodobieństwo przeżycia nie różni się pomiędzy badanymi grupami.

Tutaj kryterium wyboru, na podstawie którego można łatwo wyodrębnić dwie grupy jest płeć. Oczywiście istnieją także inne możliwości podziału, lecz ten jest jednym z najbardziej podstawowych. Test ten wykonano za pomocą funkcji *logrank_test*, która jest dostępna w programie *RStudio* po zainstalowaniu paczki *coin*. Funkcja ta zwraca wartość p, która jest kluczowa dla wyniku tego eksperymentu. Logrank test wykonano dla rozkładu wykładniczego z parametrem $\lambda = 1/10$, Weibulla z parametrami $\alpha = 1$ i $\beta = 3/2$ oraz Pareto z parametrami $k = 4$ i $\alpha = 3$.

Dla każdego z powyższych rozkładów w tabeli 2.2 przedstawiono odpowiadającą mu wartość p. W analizowanym tutaj przypadku poziom istotności $\alpha = 0.05$. Przypomnijmy, że hipoteza zerowa brzmi: *Ryzyko śmierci jest takie samo w obu grupach.* Tutaj grupy wyodrębniono ze względu na płeć; została ona również przyporządkowana syntetycznie za pomocą losowania – każda z nich z prawdopodobieństwem 1/2.

Źródło: Opracowanie własne

Tabela 2.2: Prawdopodobieństwo testowe

	wartość p
Rozkład wykładniczy	0.936
Rozkład Weibulla	0.03057
Rozkład Pareto	0.7113

Aby lepiej zrozumieć działanie tego typu testu, należy najpierw zastanowić się czym jest wartość p (prawdopodonieństwo testowe). Często można spotkać się również z nazewnictwem anglojęzycznym – wtedy wartość ta określana jest jako *p-value*.

Definicja 2.5 (Wartość p)**.** Wartość prawdopodobieństwa, która sprzyja uzyskaniu danego rezultatu, jeśli hipoteza zerowa jest prawdziwa jest nazywana *wartością p*. [8]

Alternatywna definicja określa tę wartość jako najmniejszy poziom ufności, przy którym odrzucamy hipotezę zerową na podstawie zaobserwowanej wartości statystyki testowej. W skrócie, wartość p decyduje o tym, czy hipoteza zerowa zostanie odrzucona. Jeśli $p \leqslant \alpha$ to należy odrzucić H_0, zaś gdy $p > \alpha$ to nie ma podstaw do jej odrzucenia.

Zatem w przypadku rozkładu wykładniczego oraz Pareto nie ma podstaw do odrzucenia hipotezy, zaś w przypadku Weibulla należy ją odrzucić.

2.5 Metody parametryczne

W naukach statystycznych badających śmiertelność populacji poza metodami nieparametrycznymi takimi jak estymator Kaplana-Meiera czy też logrank test, wyróżnia się także metody parametryczne. Jedno z podejść zakłada stworzenie modelu dla wskaźnika śmiertelności i określenie rozkładu prawdopodobieństwa czasu przeżycia.

Jak zwykle podczas takiego typu eksperymentów w celu usystematyzowania działania tej metody wprowadzono kilka oznaczeń:

- $\lambda(t)$ (j.ang. *hazard function*) – funkcja wskaźnika śmietelności od czasu,
- $f(t)$ – gęstość prawdopodobieństwa czasu przeżycia,
- $F(t)$ – dystrybuanta dla rozkładu prawdopodobieństwa czasu przeżycia odpowiadająca gęstości $f(t)$.

Powyższe wartości są powiązane ze sobą zależnością [1]

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = f(t)/S(t) \tag{2.13}$$

gdzie $S(t) = 1 - F(t)$ jest funkcją przeżycia. Sposób rozwiązywania równania (2.13) polega na obustronnym całkowaniu po t. Jeśli funkcja wskaźnika śmiertelności od czasu jest stałą, np. $\lambda(t) = \lambda$ dla każdego t, to

$$\lambda t = -\ln[S(t)]. \tag{2.14}$$

W wyprowadzeniu wzoru (2.14) wykorzystano fakt, że $F(t)$ jest funkcją pierwotną dla $f(t)$. Po przekształceniu mamy $S(t) = exp(-\lambda t)$. Metodą największej wiarogodności (*maximum likelihood estimation*) można wyestymować λ, przy czym należy pamiętać, że dane używane do tego celu składają się również z obserwacji cenzurowanych.

Istnieje możliwość przyjęcia innych, bardziej złożonych modeli parametrycznych. Jednym z nich jest model Weibulla dany wzorem [1]

$$\lambda(t) = \alpha\gamma t^{\gamma-1} \tag{2.15}$$

gdzie $\gamma > 1$. Jest to bardzo popularny model, ponieważ odzwierciedla on częstość zapadania na nowotwory złośliwe w zależości od wieku jednostki. Z kolei model Gompetrza [1]

$$\lambda(t) = \alpha exp(\beta t) \tag{2.16}$$

jest wykorzystywany w określaniu ryzyka zgonu dorosłych pacjentów na wiele chorób, ponieważ jego przybliżenie jest dość dobre. Ryzyko to rośnie w sposób wykładniczy wraz z wiekiem jednostki. Wadą tego modelu jest tendencja do jego spadku po upływie jakiejś ilości czasu.

Rozdział 3

Analiza danych rzeczywistych

3.1 Przygotowanie danych

Podstawowym procesem umożliwiającym dostrzeżenie zmian, cech czy właściwości jakiegoś obiektu jest obserwacja. Oczywistym jest, że jej wyniki chcemy zachować w jakikolwiek ale też możliwie jak najbardziej precyzyjny sposób. W większości przypadków zapis umożliwiają wartości liczbowe. Te zabrane za pomocą doświadczenia czy też eksperymentu nazywane są danymi empirycznymi. Nie jest ważny tylko fakt ich zebrania ale także ich jakość statystyczna. Warto pamiętać, że istnieje wiele możliwości nieprawidłowości. Może być nią przykładowo pomyłka przy uzupełnianiu rubryk przez osobę za to odpowiedzialną, braki pojedynczych rekordów lub występowanie wartości niemożliwych do osiągnięcia w normalnych warunkach. Przykładowo gdyby w zestawie danych znaleziono wartość BMI wynoszącą 60, z pewnością poddano by w wątpliwość rzetelność takiej obserwacji.

Istnieją liczne sposoby rozwiązywania problemów, które zostały powyżej wymienione i nierzadko pojawiają się w trakcie analizy danych rzeczywistych. Gdy do badanego przypisanych jest kilka zmiennych istnieje ryzyko luki w każdej z nich. Jeśli ilość zmiennych nie jest duża, przykładowo brak wartości jednej zmiennej gdy ogółem jest ich znacznie więcej, wystarczy zastąpić rekord ten średnią arytmetyczną wyliczoną z wartości osiągniętych dla całej listy badanych. Gdy mamy do czynienia z obszernymi bazami danych nie wpływa to znacząco na jakość ostatecznych wyników. W przypadku, gdy w obrębie jednego wiersza można zaobserwować wiele luk wtedy należy go usunąć, ponieważ w takim wypadku lepszym rozwiązaniem jest strata niewielkiej ilości danych niż syntetyczne uzyskiwanie dużej ich części.

Ważnym narzędziem statystycznym wykorzystywanym w analizie przeżycia, które zostało opisane w poprzednim rozdziale jest estymator Kaplana-Meiera. Zastosowano go dla różnych rozkładów prawdopodobieństwa, co oznacza, że użyte wartości miały charakter czysto teoretyczny. W przypadku tworzenia takich symulacji warto wykonywać analizy także dla danych empirycznych, aby przetestować skuteczność danej metody statystycznej.

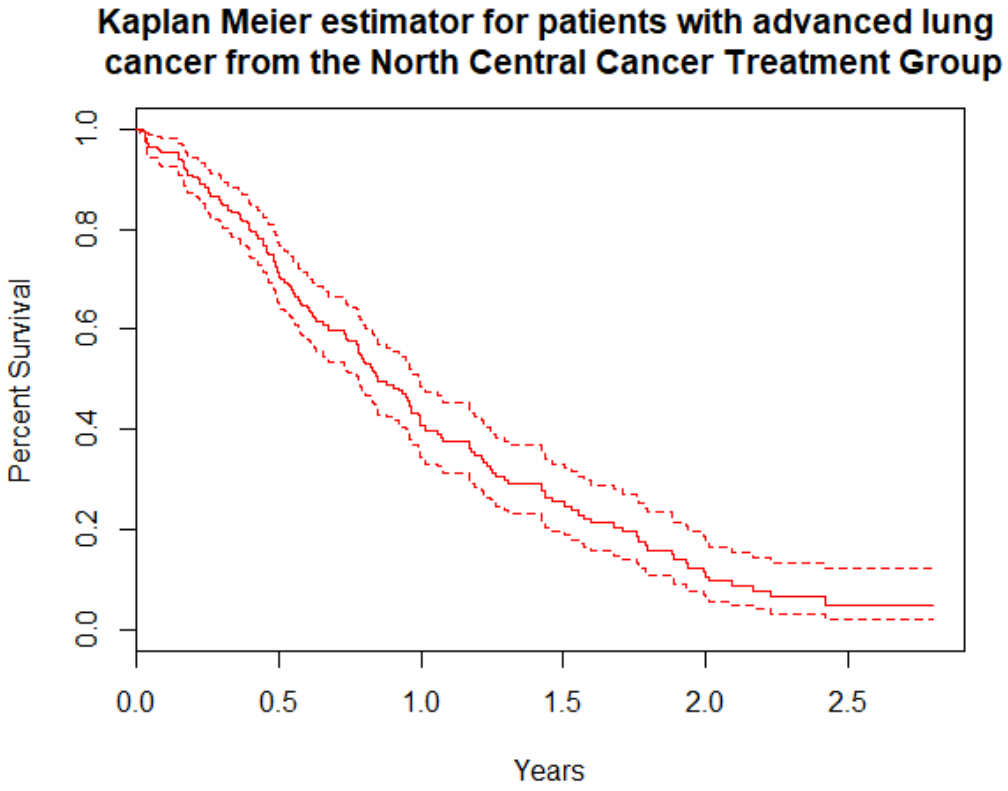
Dane empiryczne, które zostały wykorzystane do estymacji pozyskałam z pakietu *survival*; znajduje się on w dość popularnym programie statystycznym o nazwie *RStudio*. Dotyczą one chorych na różnorakie schorzenia, które w większości przypadków prowadzą do śmierci pacjenta.

3.2 Analizy dla danych empirycznych – estymator Kaplana-Meiera

W programie *RStudio* można znaleźć szeroki wachlarz zbiorów danych, na których istnieje możliwość przeprowadzenia różnych analiz statystycznych. W kręgu mojego zainteresowania znalazły się dane dotyczące pacjentów chorujących na nowotwory płuc i jelita grubego, a także na gammapatię monoklonalną o nieokreślonym znaczeniu (MGUS). Dzięki temu, że skorzystano z możliwości użycia gotowych danych, nie ma konieczności nanoszenia jakichkolwiek zmian. Gdyby były one pozyskane w tradycyjny sposób (uzupełnianie tabeli na podstawie wyników dla każdego pacjenta), wtedy najpierw należałoby przyjrzeć się wykorzystywanym wynikom; analizy wykonane na nieprawidłowych danych nie mają żadnej wartości statystycznej.

W pierwszej kolejności estymowanie przeprowadzono dla chorych na nowotwór złośliwy płuc. Przeżywalność nie jest duża i waha się w zależności od jego rodzaju. Dokładniejsze obserwacje można odczytać z rysunku 3.1.

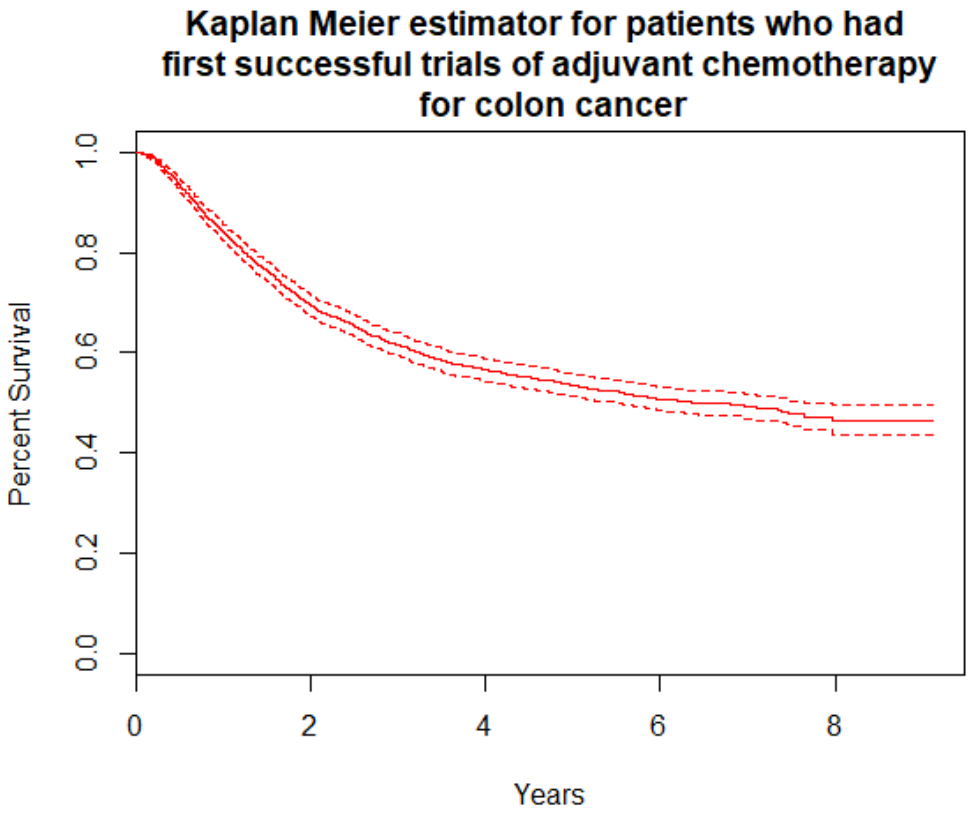
Źródło: Opracowanie własne



Rysunek 3.1: Estymator Kaplana-Meiera dla pacjentów z grupy klinicznej *North Central Cancer Treatment* chorych na zaawansowany nowotwór płuc

Prawdopodobieństwo przeżycia jest bliskie zeru już po upływie 2.5 roku. Sytuacja wygląda inaczej w przypadku chorych na nowotwór jelita grubego leczonych chemioterapią wspomagającą. Na podstawie rysunku 3.2 można sądzić, że w wypadku zachorowania istnieje aż 50% szans na to, że badany przeżyje ponad 8 lat od momentu postawienia diagnozy.

Źródło: Opracowanie własne



Rysunek 3.2: Estymator Kaplana-Meiera dla pierwszych pacjentów leczonych chemioterapią uzupełniającą z powodu nowotworu jelita grubego

Gammapatia monoklonalna o niezidentyfikowanym znaczeniu (MGUS) jest to choroba, która często prowadzi do rozwoju szpiczaka mnogiego. Organizm chorych na MGUS wytwarza nieprawidłowe białka nazywane paraproteinami. Rokowania jednak napawają optymizmem – aż 40% badanych żyje ok. 15 lat od momentu postawienia diagnozy (wykres 3.3).

3.3 Analizy dla danych empirycznych – logrank test

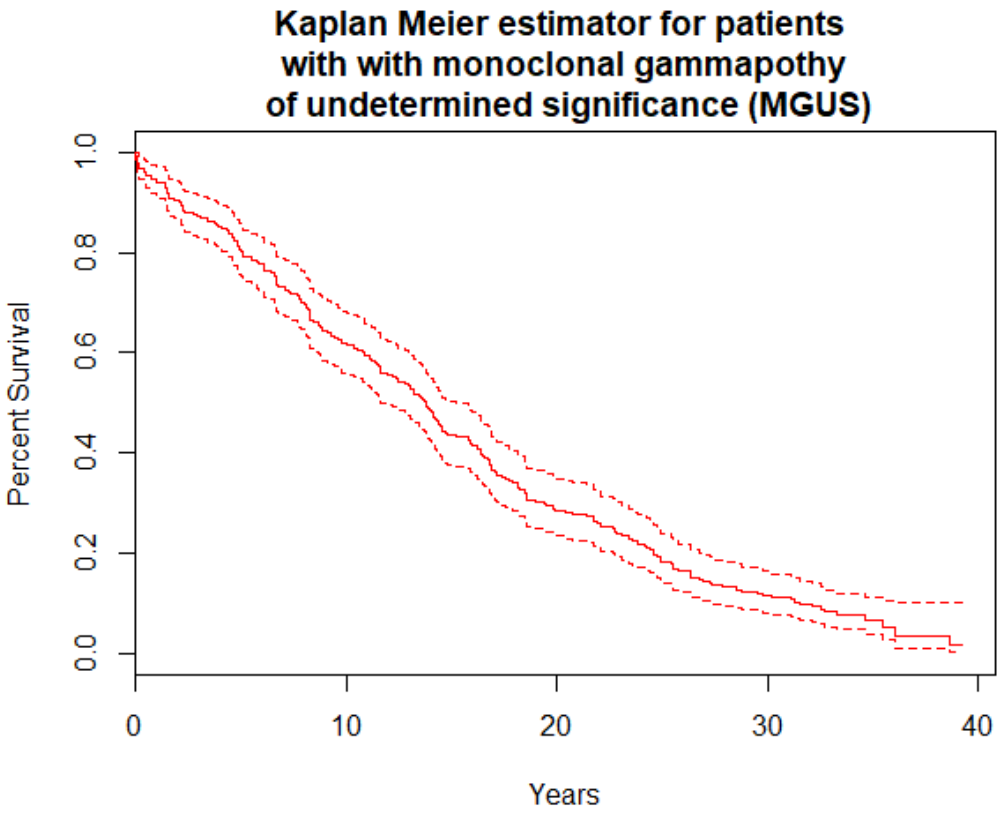
Jak wynika z przeprowadzonych dotychczas badań i zebranych danych medycznych, nowotwór złośliwy płuc dotyka znacznie częściej mężczyzn niż kobiety. Analogiczne zjawisko można zaobserwować w przypadku gamapatii monoklonalnej o niezidenty-

Źródło: Opracowanie własne

Tabela 3.1: Prawdopodobieństwo testowe

	wartość p
Nowotwór płuc	0.001
Nowotwór jelita grubego	0.611
Gammapatia monoklonalna o niezidentyfikowanym znaczeniu	0.01168

Źródło: Opracowanie własne



Rysunek 3.3: Estymator Kaplana-Meiera dla pacjentów chorych na gammapatię monoklonalną o niezidentyfikowanym znaczeniu (MGUS)

fikowanym znaczeniu. Co więcej, analizy statystyczne również skłaniają się ku postawieniu takiej hipotezy.

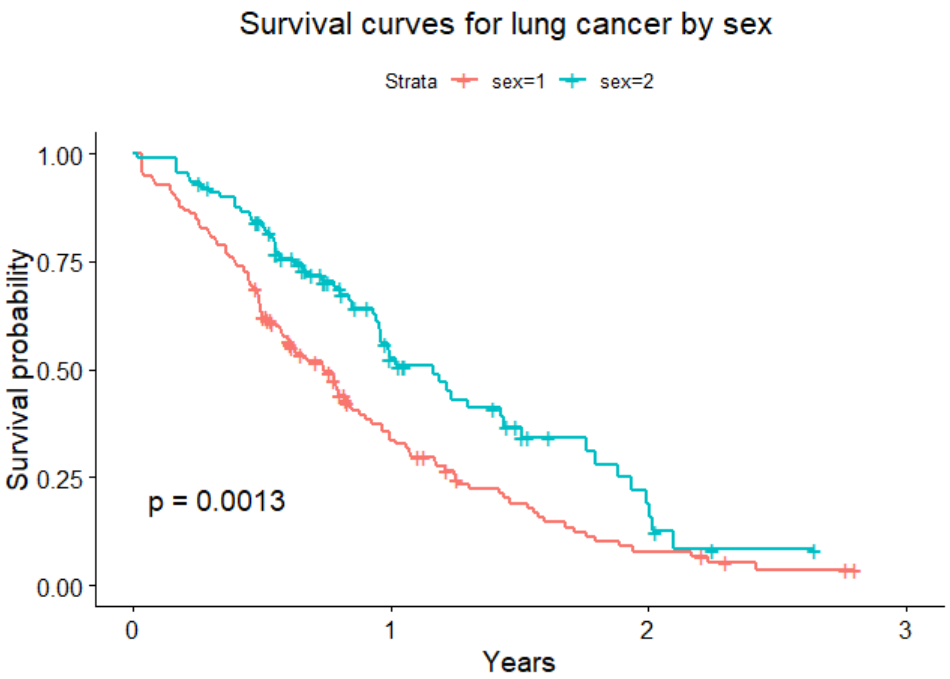
Podczas przeprowadzania testu ważne jest ustalenie poziomu istotności α. Tutaj będzie on wynosił 0.05; takie jest też domyślne ustawienie programu *RStudio*. Wartość p może należeć do dwóch przedziałów, które są wyznaczane poprzez wartość poziomu istotności.

Przypomnijmy, że dla testu logrank hipoteza zerowa brzmi: *Ryzyko śmierci jest takie samo w obu grupach*. W przypadku tych dwóch wymienionych na początku podrozdziału chorób wartość p jest bardzo mała i ostro mniejsza od poziomu istotności α co wymusza odrzucenie tej hipotezy. Oznacza to więc, że płeć ma wpływ na prawdopodobieństwo zgonu. Sytuacja wygląda inaczej w przypadku nowotworu złośliwego jelita grubego, ponieważ wartość p jest bardzo duża i wynosi 0.611 co implikuje brak podstaw do odrzucenia hipotezy. Wyniki potrzebne do przeprowadzonych powyżej analiz zebrano w tabeli 3.1.

Dla jeszcze lepszego zobrazowania sytuacji dodatkowo wyestymujemy krzywe przeżycia przy podziale na grupy. Dla wykorzystywanych tutaj danych najlepszym i najbardziej klarownym kryterium wyboru jest płeć.

Wykres dla chorych na nowotwór złośliwy płuc przedstawiono na rysunku 3.4. Cyfrą 1 oznaczono grupę mężczyzn, zaś cyfrą 2 – kobiet. Można zauważyć, że szybkość spadku wartości funkcji przeżycia dla kobiet jest wolniejsza. Ponadto osiąga ona większe wartości. Jest to konsekwencją wyższej przeżywalności w grupie przez nie wyodrębnionej i zgadza się z powszechnie dostępną obecnie wiedzą medyczną – mężczyźni bowiem w wyniku tej

Źródło: Opracowanie własne



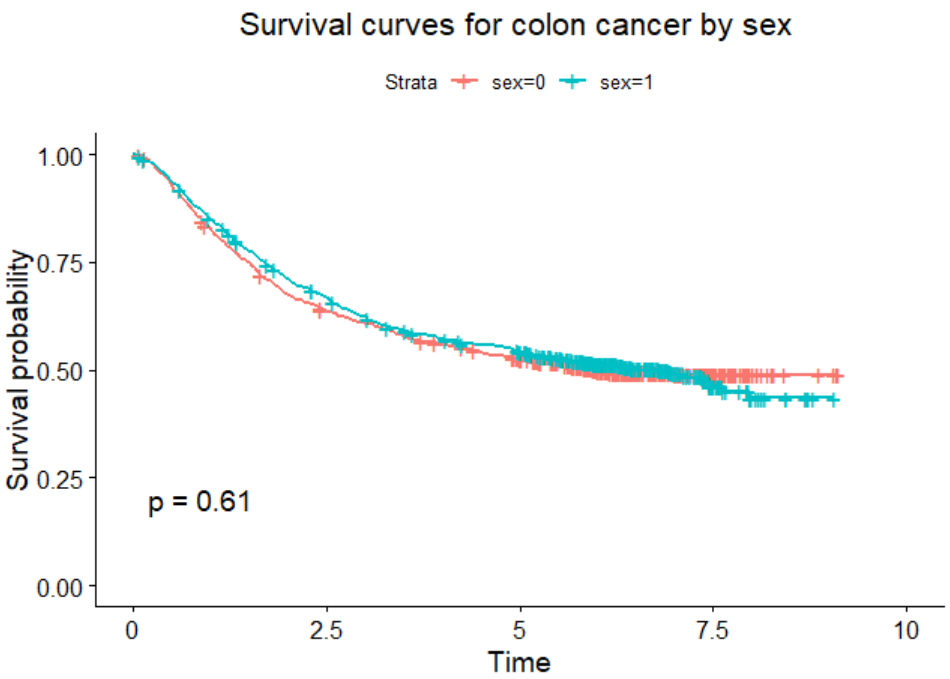
Rysunek 3.4: Krzywe przeżycia dla pacjentów chorych na nowotwór złośliwy płuc – podział ze względu na płeć

choroby umierają zdecydowanie częściej.

Na wykresie 3.5 widać rezultaty analogicznej analizy dla chorych na nowotwór złośliwy jelita grubego leczonych chemioterapią wspomagającą. Nie widać tam znaczącego wpływu płci na prawdopodobieństwo zgonu. Taki wniosek nie powinien dziwić, ponieważ wynik testu logrank na poziomie większym od α implikował brak podstaw do odrzucenia hipotezy na temat takiego samego ryzyka śmierci w obu grupach.

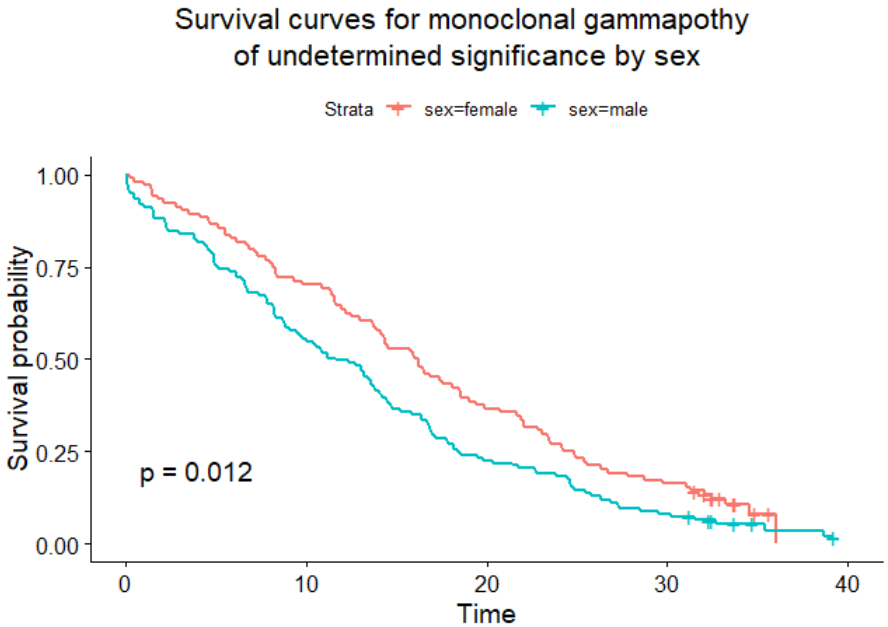
Ostatni rysunek dotyczy chorych na gammapatię monoklonalną o niezidentyfikowanym znaczeniu. Nie jest to schorzenie szybko postępujące. Prawdopodobieństwo przeżycia wynosi zero dopiero po około 40 latach od momentu zdiagnozowania danej jednostki. Jednak w przypadku, gdy zostaną wyestymowane funkcje przeżycia z podziałem ze względu na płeć, jest widoczna wyraźna rozbieżność (rysunek 3.6). Tutaj kolor niebieski oznacza mężczyzn, a czerwony – kobiety. Okazuje się, że prognoza przeżycia jest różna dla każdej z grup. Linia koloru czerwonego jest wyraźnie powyżej linii koloru niebieskiego, co oznacza, że ponownie prawdopodobieństwo przeżycia w przypadku kobiet jest wyższe. Ponadto ta tendencja utrzymuje się przez cały okres trwania choroby.

Źródło: Opracowanie własne



Rysunek 3.5: Krzywe przeżycia dla pacjentów chorych na nowotwór złośliwy jelita grubego leczonych chemioterapią wspomagającą – podział ze względu na płeć

Źródło: Opracowanie własne



Rysunek 3.6: Krzywe przeżycia dla pacjentów chorych na gammapatię monoklonalną o niezidentyfikowanym znaczeniu – podział ze względu na płeć

Podsumowanie

Metody analizy przeżycia pozwalają na obserwację jednostek danej populacji z bardzo szerokiej perspektywy. Można powiedzieć, że w pewien sposób systematyzują dane medyczne, ponieważ dzięki analizom pozyskuje się różnego rodzaju wnioski i przedstawia je w zwartej formie – wykresów, tabel czy obliczonych wartości wskaźników.

Tablice trwania życia powstały na początku prób oszacowania prawdopodobieństwa przeżycia – z tego też powodu jest to metoda dość przestarzała i obecnie rzadko wykorzystywana. Narzuca także pewne ograniczenia – przykładowo w pokoleniowej tablicy przeżycia badani muszą być urodzeni w danym przedziale czasowym.

Zdecydowanie lepszą metodą jest estymator Kaplana-Meiera. Daje on statystykom obraz funkcji przeżycia, która pozwala przewidzieć prawdopodobieństwo śmierci jednostki. Ponadto w doskonały sposób pozwala na wykorzystywanie obserwacji uciętych, ponieważ prawdopodobieństwo przeżycia jest liczone za pomocą iloczynu prawdopodobieństw dla kolejnych przedziałów czasowych. Myślę, że to dobrze rokujące na przyszłość narzędzie statystyczne; prawdopodobieństwo przeżycia wydaje się być bardzo interesującym parametrem dla naukowców różnych dziedzin. Co więcej, jeszcze nie tak dawno (około 60 lat temu), żaden statystyk nie miał go do dyspozycji, co niewątpliwie było bardzo ograniczające dla wyników eksperymentów. Ponadto estymator ten stwarza możliwość prognozowania funkcji przeżycia w grupach.

W analizie przeżycia wykorzystywane są również metody parametryczne, które mają na celu dopasowanie rozkładu czasu trwania życia do znanych rozkładów prawdopodobieństwa, m. in. wykładniczego czy też Weibulla, co zostało dokładniej opisane w jednym z podrozdziałów. Niestety wadą metod parametrycznych jest czasem problem z wybraniem odpowiedniego modelu; istnieją przypadki w których nie istnieje możliwość podparcia się żadną teorią.

Test logrank jest testem nieparametrycznym. Czasem nazywany jest testem Mantela-Coxa. Wykonuje się go przy podziale na grupy ze względu na ustalone kryterium wyboru. Test ten porównuje funkcje przeżycia w tych grupach. Dzięki temu można stwierdzić, czy ryzyko śmierci jest takie samo w obrębie każdej z nich. Przeprowadzone za jego pomocą analizy prawidłowo doprowadziły do wniosku, że w przypadku nowotworu złośliwego płuc oraz gammapatii monoklonalnej o niezidentyfikowanym znaczeniu większe ryzyko śmierci ma miejsce w grupie mężczyzn.

W części badawczej pracy sprawdzono działanie tych metod dla danych teoretycznych pochodzących z rozkładów prawdopodobieństwa oraz dla danych empirycznych pozyskanych dzięki obserwacjom populacji chorych. Łatwo zauważyć, że działają one w pożądany sposób. Estymator Kaplana-Meiera z dużą dokładnością prognozuje funkcję przeżycia, zaś logrank test prawidłowo ocenia ryzyko śmierci spowodowanej przez wybraną chorobę przy podziale na grupy.

Bibliografia

[1] P. Armitage, G. Berry, J.N.S. Matthews: *Statistical Methods in Medical Research*, Blackwell Science, rozdział 17, 2002r.

[2] Nora Pashayan, Steve Morris, Fiona J. Gilbert, Paul D. P. Pharoah: *Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer (A Life-Table Model)*, JAMA Oncology, 2018r. `https://europepmc.org/articles/PMC6230256`

[3] E. L. Lehmann, G. Casella, S. Fienberg, I. Olkin: *Testing Statistical Hipotheses*, Springer Texts in Statistics, 1986r.

[4] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore: *Understanding survival analysis: Kaplan-Meier estimate*, International Journal of Ayurveda Research, 2010r.

[5] David G. Kleinbaum: *Kaplan-Meier Survival Curves and the Log-Rank Test*, Springer, 1996r.

[6] Mark Stevenson: *An Introduction to Survival Analysis*, EpiCentre, IVABS, Massey University, 2007r.

[7] Smita Skrivanek: *Power of Statistical Test*, MoreSteam, LLC, 2009r. `https://media.moresteam.com/main/downloads/power-stat-test.pdf`

[8] John H. Mcdonald: *Handbook of biological statistics*, Sparky House Publishing, University of Delaware, Baltimore, Maryland, U.S.A., 2014r.

[9] David S. Moore, George P. McCabe, Bruce A. Craig: *Introduction to the Practice of Statistics*, W. H. Freeman and Company New York, Purdue University, 2009r.