

Ewa Wycinka

Modele zdarzeń konkurujących i ich zastosowania w ocenie ryzyka niewypłacalności pożyczkobiorcy

Wydawnictwo Uniwersytetu Gdańskiego

Modele zdarzeń konkurujących
i ich zastosowania w ocenie ryzyka
niewypłacalności pożyczkobiorcy

Ewa Wycinka

Modele zdarzeń konkurujących i ich zastosowania w ocenie ryzyka niewypłacalności pożyczkobiorcy

Wydawnictwo Uniwersytetu Gdańskiego
Gdańsk 2019

Recenzent

Dr hab. Joanna Landmesser

Redakcja

Agnieszka Kołwzan

Projekt okładki i stron tytułowych

Karolina Johnson

Skład

Maksymilian Biniakiewicz

Książka dofinansowana z działalności statutowej

Wydziału Zarządzania Uniwersytetu Gdańskiego

© Copyright by Uniwersytet Gdański, Wydawnictwo Uniwersytetu Gdańskiego

ISBN 978-83-7865-826-9

Wydawnictwo Uniwersytetu Gdańskiego
ul. Armii Krajowej 119/121, 81-824 Sopot
tel./fax 58 523 11 37, tel. 725 991 206
e-mail: wydawnictwo@ug.edu.pl
www.wyd.ug.edu.pl

Księgarnia internetowa: www.kiw.ug.edu.pl

Druk i oprawa

Zakład Poligrafii Uniwersytetu Gdańskiego
ul. Armii Krajowej 119/121, 81-824 Sopot
tel. 58 523 14 49, fax 58 551 0532

Spis treści

Wstęp	7
Rozdział 1. Probabilistyczne ujęcie czasu	11
1.1. Czas trwania do wystąpienia pojedynczego zdarzenia	11
1.1.1. Czas trwania jako ciągła zmienna losowa	11
1.1.2. Czas trwania jako dyskretna zmienna losowa	16
1.2. Cenzurowanie: rodzaje, modele cenzurowania	17
1.3. Funkcja wiarygodności dla danych cenzurowanych	20
1.4. Czas trwania do wystąpienia pierwszego ze zdarzeń konkurujących	22
1.4.1. Teoria zdarzeń konkurujących i jej rozwój	22
1.4.2. Podejście I. Czas do wystąpienia pierwszego ze zdarzeń konkurujących jako dwuwymiarowa zmienna mieszana	25
1.4.3. Podejście II. Wielowymiarowy rozkład trwania zmiennej ukrytej	30
1.4.4. Podejście III. Zdarzenia konkurujące jako modele wielostanowe	33
Rozdział 2. Modele czasu trwania do wystąpienia pojedynczego zdarzenia	35
2.1. Nieparametryczne modele czasu trwania do wystąpienia pojedynczego zdarzenia	35
2.2. Testy jednorodności rozkładów czasu trwania w przypadku pojedynczych zdarzeń	42
2.3. Semiparametryczny model proporcjonalnego hazardu Coxa w przypadku pojedynczego zdarzenia	45
2.3.1. Postać modelu proporcjonalnego hazardu	45
2.3.2. Zmienne objaśniające i ich funkcje	46
2.3.3. Estymacja parametrów modelu	53
2.3.4. Ocena poprawności modelu	57
2.3.4.1. Reszty w modelu Coxa	57
2.3.4.2. Weryfikacja założenia proporcjonalności modelu	59
2.3.4.3. Badanie obserwacji odstających	60
2.3.5. Ocena dopasowania modelu	62
2.3.6. Dobór zmiennych do modelu	67
2.4. Model proporcjonalnego hazardu Coxa dla czasu dyskretnego	71
2.5. Model regresji zdarzeń konkurujących oparty na pseudoobserwacjach	74
2.5.1. Idea pseudoobserwacji	74
2.5.2. Pseudoobserwacje dla funkcji trwania do wystąpienia pojedynczego zdarzenia	75
2.5.3. Uogólnione równania estymacyjne pseudoobserwacji dla funkcji trwania	82

Rozdział 3. Modele czasu trwania do wystąpienia pierwszego ze zdarzeń konkurujących	86
3.1. Nieparametryczne modele czasu trwania do wystąpienia pierwszego ze zdarzeń konkurujących	86
3.2. Testy jednorodności subrozkładów	91
3.3. Modele regresji dla hazardu według przyczyn	94
3.4. Model regresji funkcji hazardu subrozkładu	99
3.5. Porównanie modelu hazardu według przyczyn i modelu hazardu subrozkładu	101
3.6. Mieszanka modeli – podejście horyzontalne	106
3.7. Mieszanka modeli – podejście wertykalne	112
3.8. Modele regresji dla subdystrybuanty oparte na pseudoobserwacjach	115
Rozdział 4. Ryzyko niewypłacalności pożyczkobiorcy i ryzyko wcześniejszej spłaty jako zdarzenia konkurujące	124
4.1. Pożyczki społecznościowe i ryzyko niewypłacalności pożyczkobiorcy	124
4.2. Ryzyko kredytowe i ryzyko niewypłacalności a ryzyko wcześniejszej spłaty	126
4.3. Główne kierunki badań nad ryzykiem niewypłacalności. Zastosowanie analizy przetrwania do badania ryzyka niewypłacalności	128
4.4. Przegląd badań ryzyka niewypłacalności pożyczkobiorców pożyczek społecznościowych	131
4.5. Platforma pożyczek społecznościowych Lending Club i pożyczki udzielone przez Lending Club w roku 2014 – charakterystyka próby badawczej	134
4.6. Prawdopodobieństwo niewypłacalności oraz prawdopodobieństwo wcześniejszej spłaty oszacowane metodami nieparametrycznymi	143
Rozdział 5. Przygotowanie zmiennych objaśniających do budowy semiparametrycznych modeli zdarzeń konkurujących	150
5.1. Przegląd metod transformacji zmiennych stosowanych w modelach przetrwania dla ryzyka kredytowego	150
5.2. Rekategoryzacja jakościowych zmiennych objaśniających	151
5.3. Poszukiwanie postaci funkcji ilościowych zmiennych objaśniających	164
5.4. Podsumowanie	171
Rozdział 6. Modele semiparametryczne ryzyka niewypłacalności przy uwzględnieniu ryzyka wcześniejszej spłaty jako zdarzenia konkurującego	174
6.1. Model regresji dla hazardu według przyczyn	174
6.2. Modele hazardu subrozkładu	186
6.3. Mieszanka modeli – podejście horyzontalne	188
6.4. Mieszanka modeli – podejście wertykalne	190
6.5. Model subdystrybuanty zbudowany dla pseudoobserwacji	194
6.6. Podsumowanie i kierunki dalszych badań	195
Podsumowanie	199
Załączniki	204
Literatura	250
Spis schematów, tabel i wykresów	262

Wstęp

Analiza trwania, znana również pod nazwą analizy przeżycia bądź analizy historii zdarzeń, stanowi grupę metod służących badaniu rozkładu zjawisk w czasie. Podstawowym zadaniem metod analizy trwania jest modelowanie procesów trwania między dwoma momentami – początkowym i końcowym. Metody te coraz szerzej wykorzystuje się w ekonomii i naukach społecznych, czego efektem są liczne monografie i artykuły naukowe zarówno o charakterze metodycznym (m.in. Frątczak, Gach-Ciepiela, Babiker 2005; Balicki 2006; Landmesser 2013), jak i aplikacyjnym¹. Frątczak, Gach-Ciepiela, Babiker (2005) dokonali następującej klasyfikacji metod analizy trwania:

- metody pojedynczych epizodów, w tym metody:
 - jednego momentu początkowego i jednego momentu końcowego²;
 - jednego momentu początkowego i wielu różnych momentów końcowych (ryzyko konkurencyjne);
- metody wielu epizodów z wieloma momentami początkowymi i wieloma momentami końcowymi.

W każdej z powyższych grup metod mogą być estymowane modele nieparametryczne, parametryczne i semiparametryczne. Inną klasyfikację przyjęła Landmesser (2013), która wyróżniła wśród metod analizy trwania modele jednowymiarowych czasów trwania oraz wielowymiarowych czasów trwania, a wśród tych ostatnich – modele wielokrotnych epizodów. Przyjmowane w literaturze nazewnictwo dotyczące metod analizy trwania, jak i klasyfikacja tych metod są więc zróżnicowane. Analizowane w niniejszej pracy modele można zaklasyfikować do modeli pojedynczych epizodów, według klasyfikacji Frątczak, Gach-Ciepiela i Babikera (2005). Modele te zostały podzielone przez autorkę na modele czasu trwania do wystąpienia pojedynczego zdarzenia (jeden moment początkowy i jeden rodzaj zdarzenia określający moment końcowy) oraz modele czasu trwania do wystąpienia pierwszego ze zdarzeń konkurujących (jeden moment początkowy i wiele różnych zdarzeń określających moment końcowy). Natomiast według klasyfikacji Landmesser (2013) rozważane modele można zakwalifikować zarówno do modeli jednowymiarowych czasów trwania, jak i modeli wielowymiarowych czasów trwania (z pominięciem modeli wielokrotnych epizodów).

¹ Obszerny przegląd piśmiennictwa w zakresie wykorzystania metod analizy trwania w ekonomii przedstawiono w pracy Landmesser 2013.

² W pracy Frątczak, Gach-Ciepiela, Babiker 2005 moment początkowy jest nazywany „stanem wyjścia”, a moment końcowy „stanem przeznaczenia”.

W praktyce, gdy jest mowa o metodach analizy trwania, najczęściej rozważa się tylko przypadek wystąpienia jednego momentu początkowego i jednego rodzaju momentu końcowego. Przyjmuje się więc założenie, że każda jednostka musi doznać zdarzenia określającego wystąpienie momentu końcowego. Założenie to ogranicza przydatność metod analizy trwania w sytuacjach, gdy badane są zdarzenia, które nie realizują się u wszystkich jednostek, lub gdy rozważa się przyczyny realizacji zdarzeń, np. przy badaniu upadłości przedsiębiorstw czy niewypłacalności kredytobiorców. Jeżeli analizowane są różne typy momentów końcowych, przeważnie nie używa się ogólnego pojęcia „analiza trwania” czy „analiza przeżycia”, lecz mówi się o analizie lub teorii zdarzeń konkurujących (ang. *competing risks*). W tym kontekście teoria zdarzeń konkurujących stanowi rozwinięcie analizy czasu trwania na przypadek wystąpienia jednej z różnych przyczyn zdarzenia kończącego czas trwania jednostki.

Idea zdarzeń konkurujących została nakreślona już w pracy Bernoulliego (1766), jednak dopiero w ostatnich dwóch dekadach nastąpił wzrost zainteresowania tymi metodami i coraz częściej są podejmowane próby zastosowania modeli zdarzeń konkurujących w różnych dziedzinach. W literaturze polskojęzycznej zależne ryzyka konkurencyjne (w niniejszej monografii nazywane zdarzeniami konkurującymi) były przedstawiane w kontekście badania aktywności ekonomicznej ludności (Frątczak, Gach-Ciepiela, Babiker 1996; Landmesser 2013). Pewną lukę, zdaniem autorki, stanowi brak pracy, która w sposób kompleksowy systematyzowałaby zagadnienia dotyczące niezależnych zdarzeń konkurujących. Tylko wybrane zagadnienia związane z zastosowaniami niezależnych zdarzeń konkurujących w analizie różnych problemów ekonomiczno-społecznych podejmowane były m.in. w pracach dotyczących ryzyka niewypłacalności kredytobiorcy (Matuszyk 2015; Wycinka 2015a; 2015b; 2017a; 2017b; Wycinka, Jurkiewicz 2017; 2019) oraz w pracach odnoszących się do ryzyka wyjścia z bezrobocia (Bieszk-Stolorz 2017; 2018). Kolejnym powodem podjęcia tematu przez autorkę było to, że w badaniach ryzyka niewypłacalności wykorzystywano do tej pory głównie modele pojedynczych zdarzeń. Natomiast w pracach, w których identyfikowano wcześniejszą spłatę kredytu jako zdarzenie konkurujące (m.in. Stepanova, Thomas 2002; Andreeva 2006; Matuszyk 2015), rozważano oddzielnie rozkłady brzegowe poszczególnych czasów zdarzeń bez analiz rozkładów łącznych. Wkładem autorki w badania nad modelami czasów do wystąpienia niewypłacalności jest propozycja wykorzystania subdystrybuant rozkładów jako modeli zdarzeń konkurujących. Zagadnienie to zostało naszkicowane w pracy *Competing risks models of default in the presence of early repayment* (Wycinka 2019) oraz rozwinięte w niniejszej monografii.

Celem monografii jest wskazanie metod odpowiednich do modelowania czasów trwania do wystąpienia niezależnych zdarzeń konkurujących oraz ocena przydatności tych metod do badania ryzyka niewypłacalności pożyczkobiorcy. Prezentowane metody są uniwersalne w obszarze możliwych ich zastosowań. W niniejszej pracy metody te przedstawiono w sposób ogólny tak, aby umożliwiło to łatwą ich aplikację do rozwiązywania różnych problemów badawczych z zakresu ekonomii i nauk społecznych.

Zostały sformułowane następujące cele szczegółowe o charakterze poznawczym:

1. identyfikacja funkcji opisujących rozkład zdarzeń w przypadku pojedynczych i konkurujących zdarzeń,
2. systematyzacja i analiza porównawcza nieparametrycznych i semiparametrycznych modeli czasu trwania do wystąpienia pierwszego ze zdarzeń konkurujących,
3. wykrycie prawidłowości statystycznych w kształtowaniu się ryzyka niewypłacalności w pożyczkach długoterminowych.

Cele praktyczne pracy polegały na:

1. skonstruowaniu modeli trwania ze względu na ryzyko niewypłacalności dla pożyczek długoterminowych,
2. empirycznej weryfikacji przydatności zaprezentowanych metod zdarzeń konkurujących w ocenie niewypłacalności pożyczkobiorców w przypadku pożyczkobiorców długoterminowych,
3. wskazaniu zagrożeń przy stosowaniu metod analizy trwania zarówno dla zdarzeń pojedynczych, jak i konkurujących.

W pracy weryfikowane są następujące hipotezy badawcze:

1. ryzyko niewypłacalności i ryzyko wcześniejszej spłaty są zdarzeniami konkurującymi i konstrukcja modeli czasu trwania do wystąpienia niewypłacalności wymaga uwzględnienia rozkładu czasu do wystąpienia wcześniejszej spłaty,
2. połączenie różnych metod transformacji zmiennych objaśniających pozwala na wykrycie rodzaju związku łączącego poszczególne zmienne objaśniające ze zmienną objaśnianą w semiparametrycznych modelach zdarzeń konkurujących,
3. modele niewypłacalności w długoterminowych umowach są wrażliwe na występowanie wartości wpływowych nawet w przypadku bardzo dużych prób,
4. stosowanie różnych metod weryfikacji założenia proporcjonalności hazardów może dawać różne wyniki, co nie jest obojętne dla poprawności i jakości budowanych modeli.

Praca składa się z części teoretycznej (rozdziały 1–3) oraz części empirycznej (rozdziały 4–6). W części teoretycznej przedstawiono podstawowe funkcje opisujące rozkład trwania w przypadku rozważania pojedynczych zdarzeń, jak i zdarzeń konkurujących. Nakreślono również najważniejsze etapy rozwoju teorii zdarzeń konkurujących. Wprowadzony w pierwszym rozdziale aparat pojęciowy jest wykorzystany w rozdziale drugim, w którym scharakteryzowano nieparametryczne modele czasu trwania oraz semiparametryczny model proporcjonalnych hazardów w przypadku pojedynczego zdarzenia. Szczególną uwagę poświęcono metodom badania poprawności i jakości modelu oraz weryfikacji założenia proporcjonalności. Jako alternatywne podejście przedstawiono modele oparte o pseudoobserwacje, niewymagające przyjmowania założenia proporcjonalności hazardów. Na bazie modeli dla zdarzenia pojedynczego zaprezentowanych w rozdziale drugim konstruowane są modele zdarzeń konkurujących, których systematykę zawiera rozdział trzeci.

W rozdziale czwartym, poprzez analogię do ryzyka kredytowego, nakreślono istotę badania niewypłacalności pożyczkobiorcy. Badanie empiryczne przeprowadzono na próbie pożyczkobiorców największej platformy pożyczek społecznościowych w USA. Dane te są jawne i w pełni dostępne, co umożliwia weryfikację wyników

badani i ich porównywalność z badaniami innych autorów. W rozdziale czwartym dokonano przeglądu badań naukowych w zakresie modeli ryzyka niewypłacalności, które były prowadzone na tej samej populacji. Wyodrębniono najważniejsze predyktory ryzyka niewypłacalności. Zidentyfikowano luki badawcze w postaci braku analiz pożyczkobiorców długoterminowych (60 miesięcy) oraz pomijania w badaniach kwestii wyboru odpowiedniej postaci funkcji zmiennych objaśniających.

Rozdział piąty został poświęcony poszukiwaniom optymalnych postaci funkcji dla ilościowych i jakościowych predyktorów niewypłacalności oraz wcześniejszej spłaty pożyczki. Na bazie tak przygotowanych zmiennych w rozdziale szóstym zbudowano modele ryzyka niewypłacalności z uwzględnieniem ryzyka wcześniejszej spłaty jako zdarzenia konkurującego. Zawarto tu również dyskusję na temat zalet i ograniczeń modeli zdarzeń konkurujących wynikających ze specyfiki ryzyka kredytowego. Szczególną uwagę poświęcono konsekwencjom niespełnienia założenia proporcjonalności hazardów w przypadku predyktorów ryzyka niewypłacalności w modelach dla umów długoterminowych.

Problem zdarzeń konkurujących występuje w badaniach wielu różnych zjawisk ekonomicznych. Modele zaprezentowane w tej monografii w kontekście ryzyka niewypłacalności pożyczkobiorców mogą być zastosowane do oceny ryzyka kredytowego. Metody te mają charakter uniwersalny i mogą być bezpośrednio przełożone również na inne problemy badawcze.

* * *

Chciałabym podziękować wszystkim osobom, bez których wsparcia nie powstałaby ta praca. W szczególności składam podziękowania dr hab. Joannie Landmesser za cenne uwagi, które wpłynęły na ostateczny kształt niniejszej monografii. Pragnę również podziękować dr hab. Beacie Jackowskiej, prof. UG, która jako pierwsza oczami matematyka spojrzała na powstającą publikację oraz dr Tomaszowi Jurkiewiczowi – za wprowadzenie w tajniki programowania w R.

Rozdział 1

Probabilistyczne ujęcie czasu

1.1. Czas trwania do wystąpienia pojedynczego zdarzenia

W analizie przetrwania przedmiotem obserwacji jest przedział czasu między dwoma zdarzeniami: początkowym i końcowym. Zdarzenie początkowe inicjuje stan, w jakim znajdują się jednostki populacji, a zdarzenie końcowe, które dalej będzie nazywane zdarzeniem, powoduje, że jednostka wychodzi z tego stanu. Przyjmuje się założenie, że jednostka może doznać zdarzenia tylko raz i nie ma możliwości powrotu do populacji¹. Przedział czasu między zdarzeniem początkowym a końcowym jest określany czasem trwania (Rossa 2005a, s. 10). Czas trwania dla każdej jednostki jest mierzony od zdarzenia początkowego. W rezultacie w większości badań wykorzystywane są dwie skale czasu: czas kalendarzowy oraz czas badania.

Czas trwania jest nieujemną zmienną losową, oznaczaną jako T , którą opisuje się przez funkcje probabilistyczne: gęstości prawdopodobieństwa, dystrybuanty, trwania, hazardu i skumulowanego hazardu. Każda z tych funkcji w sposób jednoznaczny wyznacza rozkład zmiennej T i umożliwia wyznaczenie pozostałych funkcji (Jackowska 2013). Jak podkreślają Hosmer, Lemeshow, May (2008, s. 16), specyfiką analizy przetrwania jest to, że kluczową rolę odgrywa w niej estymacja dystrybuanty rozkładu, zaś drugorzędą – parametry rozkładu i ich estymacja. Zmienna losowa T to zmienna ciągła, jednak ze względu na ograniczenia metod pomiaru w niektórych sytuacjach zasadne jest traktowanie jej jako zmiennej dyskretnej.

1.1.1. Czas trwania jako ciągła zmienna losowa

Niech n będzie wielkością próby ($i = 1, \dots, n$), $T_1, T_2, \dots, T_n$ będą nieujemnymi, niezależnymi zmiennymi losowymi o identycznym rozkładzie zdefiniowanym przez funkcję dystrybuanty F . Zmienne T_i wyznaczają czas trwania do wystąpienia badanego zdarzenia u jednostki i , a ich dystrybuanty są sobie równe:

¹ Założenie to nie jest konieczne, gdy rozważa się modele wielokrotnych epizodów, w których jednostka może wielokrotnie doznawać kolejno zdarzeń początkowych i końcowych, co oznacza, że wielokrotnie powraca do populacji. Przykładem takich populacji są bezrobotni, dla których podjęcie pracy jest zdarzeniem powodującym opuszczenie populacji bezrobotnych, a utrata pracy może być zdarzeniem powodującym powrót do populacji bezrobotnych (szerzej modele te są opisane m.in. w pracy Landmesser 2013).

$$F_1 = F_2 = \dots = F_n = F.$$

Funkcja gęstości rozkładu zmiennej losowej T (ang. *density function*) jest określona wzorem

$$f(t) = \lim_{\delta t \rightarrow 0} \frac{P(t < T \leq t + \delta t)}{\delta t}. \quad (1.1)$$

Wyrażenie $f(t) \delta t$ można rozumieć jako graniczne prawdopodobieństwo, że jednostka dozna zdarzenia w nieskończenie małym przedziale czasu $(t, t + \delta t)$. Funkcja gęstości spełnia następujące warunki:

$$\forall t \in \mathbb{R} f(t) \geq 0 \quad (1.2)$$

oraz

$$\int_0^{+\infty} f(t) dt = 1. \quad (1.3)$$

Dla $T < 0$ funkcja gęstości przyjmuje wartość równą zero.

Funkcja dystrybucyjna lub inaczej dystrybuanta (ang. *distribution function, cumulative distribution function*) określona jest wzorem

$$F(t) = P(T \leq t) \quad (1.4)$$

i wyraża prawdopodobieństwo, że jednostka doświadczy zdarzenia najpóźniej w momencie t .

Relację między funkcją gęstości a dystrybuantą można przedstawić jako

$$f(t) = \frac{\partial F(t)}{\partial t}, \quad (1.5)$$

o ile funkcja f jest ciągła w punkcie t lub jako

$$F(t) = \int_0^t f(u) du. \quad (1.6)$$

Dystrybuanta musi spełniać następujące warunki:

$$1) \lim_{t \rightarrow -\infty} F(t) = 0, \quad (1.7)$$

$$2) \lim_{t \rightarrow \infty} F(t) = 1, \quad (1.8)$$

3) F jest funkcją niemalejącą,

4) F jest funkcją co najmniej prawostronnie ciągłą (Jakubowski, Sztencel 2006).

Funkcja trwania $S(t)$, nazywana również funkcją przetrwania lub funkcją dożycia (ang. *survival function*), jest zdefiniowana jako

$$S(t) = P(T > t) \quad (1.9)$$

i stanowi ona dopełnienie do jedności dystrybucyjnej rozkładu

$$S(t) = 1 - F(t). \quad (1.10)$$

Określa prawdopodobieństwo, że jednostka nie dozna zdarzenia do momentu t . Czas trwania jednostki jest więc dłuższy niż t .

Funkcja trwania spełnia następujące warunki:

$$1) \lim_{t \rightarrow -\infty} S(t) = 1, \quad (1.11)$$

$$2) \lim_{t \rightarrow \infty} S(t) = 0, \quad (1.12)$$

3) S jest funkcją nierosnącą,

4) S jest funkcją co najmniej prawostronnie ciągłą.

Warunek 2) jest równoważny przyjęciu założenia, że wszystkie jednostki muszą doznać zdarzenia, nieznany jest tylko czas jego nastąpienia. Założenie takie ma uzasadnienie, gdy badane są populacje istot żywych, a analizowane zdarzenie to zgon. Natomiast w wielu innych zastosowaniach analizy przetrwania, dla większości rodzajów analizowanych zdarzeń, założenie to nie jest spełnione. Przykładowo analizowana w części drugiej niniejszej monografii zmienna czas trwania pożyczek nie zawsze kończy się wystąpieniem zdarzenia niewypłacalność pożyczkobiorcy. Dlatego w kolejnych rozdziałach będą rozważane uogólnienia metody analizy przetrwania na więcej typów zdarzeń końcowych.

Rozkład funkcji czasu trwania może być opisany również funkcją hazardu, nazywaną także funkcją intensywności (ang. *hazard/intensity/force function*) i wyrażoną wzorem:

$$h(t) = \lim_{\delta t \rightarrow 0} \frac{P(t < T \leq t + \delta t | T > t)}{\delta t}. \quad (1.13)$$

Funkcja hazardu jest graniczną wartością warunkowego prawdopodobieństwa zdarzenia przypadającą na jednostkę czasu w przedziale od t do $t + \delta t$, przy malejącej do zera długości przedziału i pod warunkiem, że dla jednostki nie wystąpiło zdarzenie do momentu t (Balicki 2006). Funkcja hazardu może być również wyrażona za pomocą wzoru

$$h(t) = \frac{f(t)}{S(t)}. \quad (1.14)$$

Spełnia ona następujące warunki:

$$1) h(t) \geq 0 \quad (1.15)$$

2) h może być funkcją malejącą, rosnącą, stałą lub jej monotoniczność może się zmieniać.

Skumulowana funkcja hazardu wyrażona jest wzorem

$$H(t) = \int_0^t h(u) du. \quad (1.16)$$

Funkcja hazardu skumulowanego nie ma prostej interpretacji w sensie probabilistycznym (Zhang 2017). Relacje między funkcjami przedstawiają wzory ((1.17)–(1.21)). Funkcja dystrybucyjna może być wyrażona za pomocą innych funkcji jako

$$\begin{aligned} F(t) &= 1 - S(t) = \int_0^t f(u) du = 1 - \exp \left[- \int_0^t h(u) du \right] = \\ &= 1 - \exp[-H(t)]. \end{aligned} \quad (1.17)$$

Funkcja trwania może być wyrażona następująco

$$S(t) = 1 - F(t) = \int_t^{+\infty} f(u) du = \exp \left[- \int_0^t h(u) du \right] = \exp[-H(t)]. \quad (1.18)$$

Funkcja gęstości rozkładu może być wyznaczona jako

$$f(t) = - \frac{\partial S(t)}{\partial t} = \frac{\partial F(t)}{\partial t} = h(t) \exp \left[- \int_0^t h(u) du \right] = - \frac{\partial \exp[-H(t)]}{\partial t}. \quad (1.19)$$

Funkcję hazardu można przedstawić jako

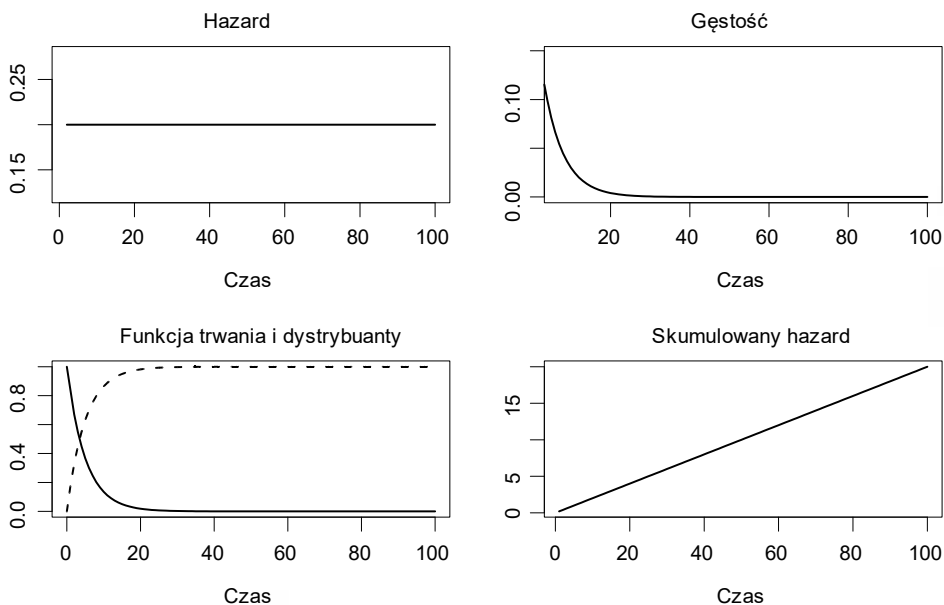
$$h(t) = -\frac{\partial \ln S(t)}{\partial t} = -\frac{\partial \ln[1-F(t)]}{\partial t} = \frac{f(t)}{\int_t^\infty f(u)du} = \frac{\partial H(t)}{\partial t}, \quad (1.20)$$

natomiast funkcję skumulowanego hazardu jako

$$H(t) = -\ln(1 - F(t)) = -\ln S(t) = \frac{\int_0^t h(u)du}{\int_u^\infty h(v)dv} = \int_0^t h(u)du \quad (1.21)$$

[Balicki 2006; Jackowska 2013; Landmesser 2013].

Graficzną prezentację zależności między pięcioma opisanymi funkcjami dla przykładowych postaci funkcji hazardu przedstawiają wykresy 1.1–1.3. Uwzględniono następujące rodzaje hazardów: stały (wykres 1.1), U-kształtny² (wykres 1.2) oraz o nieregularnym przebiegu (wykres 1.3)³. W literaturze najczęściej przedstawiane są powyższe relacje dla monotonicznej postaci funkcji hazardu, natomiast w praktyce najczęściej mamy do czynienia z funkcjami U-kształtnymi lub z funkcjami hazardu o zmiennym przebiegu monotoniczności (por. Rajarshi i Rajarshi 1988).

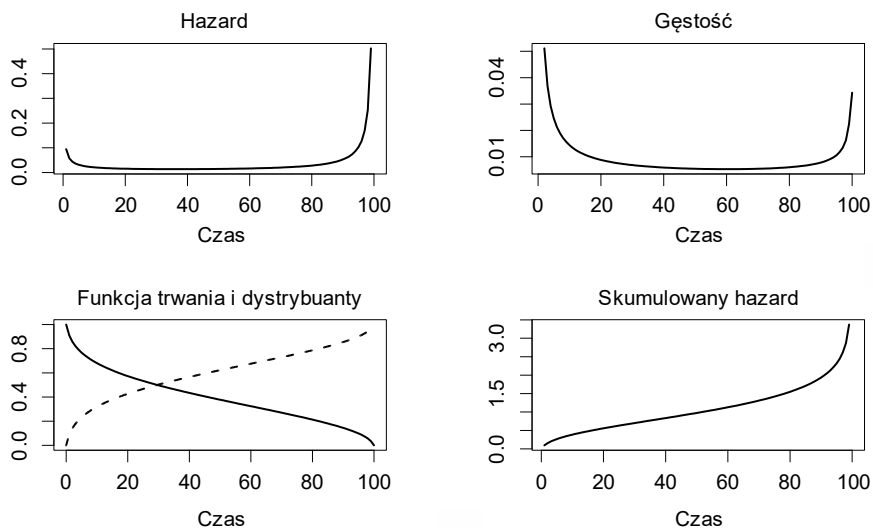


Wykres 1.1. Funkcje probabilistyczne dla rozkładu o stałym hazardzie

Źródło: opracowanie własne.

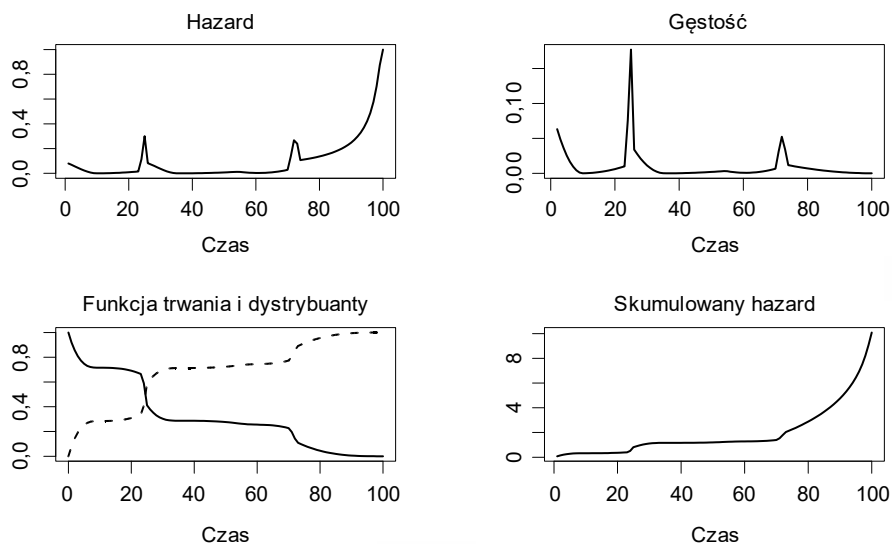
² Aby funkcja hazardu była U-kształtna, ale jednocześnie zależała od możliwie małej liczby parametrów, wykorzystano funkcję hazardu uogólnionego rozkładu jednostajnego zaproponowanego w pracy Sankaran i Jayakumar 2016.

³ Wszystkie wykresy były wyznaczane na podstawie 1000 obserwacji wygenerowanych z populacji o ustalonej (wykresy 1.1 i 1.2) funkcji hazardu lub funkcji hazardu o zmiennej monotoniczności (wykres 1.3). Obliczenia i wykresy wykonano w programie R (R Development Core Team 2017) z wykorzystaniem funkcji `sim.survdata()` z pakietu *Coxed*.



Wykres 1.2. Funkcje probabilistyczne dla rozkładu o U-kształtnym hazardzie

Źródło: opracowanie własne.



Wykres 1.3. Funkcje probabilistyczne dla rozkładu o zmiennym hazardzie

Źródło: opracowanie własne.

Czas trwania może być również opisany za pomocą prawdopodobieństw warunkowych. Warunkowe prawdopodobieństwo zdarzenia w jednostkowym przedziale czasu $(t, t + 1]$ pod warunkiem, że jednostka nie doznała zdarzenia do czasu t , wynosi

$$q_t = \frac{P(t < T \leq t+1)}{P(T > t)} = \frac{F(t+1) - F(t)}{1 - F(t)} = \frac{S(t) - S(t+1)}{S(t)}. \quad (1.22)$$

Natomiast warunkowe prawdopodobieństwo przetrwania w jednostkowym przedziale czasu $(t, t + 1]$ pod warunkiem, że jednostka nie doznała zdarzenia do czasu t , wynosi

$$p_t = \frac{P(T > t+1)}{P(T > t)} = \frac{1-F(t+1)}{1-F(t)} = \frac{S(t+1)}{S(t)}. \quad (1.23)$$

Pomiędzy warunkowym prawdopodobieństwem zdarzenia a warunkowym prawdopodobieństwem przetrwania zachodzi równość

$$p_t = 1 - q_t. \quad (1.24)$$

Zależność między funkcją trwania a warunkowymi prawdopodobieństwami zdarzenia oraz przetrwania można również wyrazić jako

$$S(t) = \prod_{k=0}^{t-1} (1 - q_k) = \prod_{k=0}^{t-1} (p_k). \quad (1.25)$$

Funkcja trwania jest więc równa iloczynowi prawdopodobieństw przetrwania kolejnych jednostkowych przedziałów czasu (Balicki 2006; Bowers i in. 1997; Jackowska 2013; Kalbfleisch, Prentice 2002; Lee, Wang 2003).

1.1.2. Czas trwania jako dyskretna zmienna losowa

Jeżeli metoda pomiaru czasu powoduje, że w jednym momencie t więcej niż jedna jednostka doznaje zdarzenia lub czas trwania jest arbitralnie podzielony na przedziały, zamiast modelu ciągłego można przyjąć model dyskretny (Rossa 2005a).

Niech T będzie dyskretną zmienną losową przyjmującą wartości t_k , ($k = 0, 1, 2, \dots$) takie, że $t_0 = 0 \leq t_1 \leq t_2 \leq t_3 \leq \dots \leq t_\omega$, gdzie ω może być skończone lub nieskończone, z prawdopodobieństwami

$$f(t_k) = P(T = t_k). \quad (1.26)$$

Funkcja trwania jest określona jako

$$S(t) = P(T > t) = \sum_{t_k > t} f(t_k), \quad (1.27)$$

a prawdopodobieństwo, że do zdarzenia dojdzie w momencie t_k , można wyrazić jako

$$f(t_k) = S(t_{k-1}) - S(t_k). \quad (1.28)$$

Funkcja hazardu dla zmiennej dyskretnej jest warunkowym prawdopodobieństwem wystąpienia zdarzenia w momencie t_k , pod warunkiem, że jednostka dotrwa do momentu t_{k-1} :

$$h(t_k) = P(T = t_k | T > t_{k-1}) = \frac{f(t_k)}{S(t_{k-1})} \quad (1.29)$$

dla $k = 1, 2, \dots, \omega$. Natomiast dla $k = 0$ zachodzi $h(t_0) = h(0) = 0$. Należy zauważyć, że w przypadku zmiennej dyskretnej wartości funkcji hazardu są ograniczone do przedziału $h(t_k) \in [0, 1]$. Na podstawie zależności określonej wzorem (1.28) można wyrazić funkcję hazardu za pomocą funkcji trwania jako

$$h(t_k) = \frac{S(t_{k-1}) - S(t_k)}{S(t_{k-1})} = 1 - \frac{S(t_k)}{S(t_{k-1})}. \quad (1.30)$$

Funkcja trwania może być wyrażona za pomocą funkcji hazardu jako

$$S(t_k) = \prod_{i=1}^k [1 - h(t_i)] \quad (1.31)$$

(Pintilie 2006, s. 26), a funkcja prawdopodobieństwa jako

$$f(t_k) = h(t_k) \prod_{i=1}^{k-1} [1 - h(t_i)]. \quad (1.32)$$

Istnieje kilka definicji funkcji hazardu skumulowanego w rozkładzie zmiennej skokowej. Według pierwszej definicji skumulowana funkcja hazardu jest równa

$$H^{(1)}(t_k) = \sum_{i=1}^k h(t_i) \quad (1.33)$$

(Pintilie 2006, s. 28). Taka funkcja skumulowanego hazardu nie spełnia jednak zależności przedstawionej wzorem (1.21). Zależność (1.21) jest podstawą dla alternatywnej definicji skumulowanego hazardu (Cox, Oakes 1984, s. 15):

$$H^{(2)}(t_k) = -\ln(S(t_k)) = -\sum_{i=1}^k \ln[1 - h(t_i)]. \quad (1.34)$$

Jeżeli funkcja hazardu przyjmuje małe wartości w punktach t_k , to funkcje hazardu skumulowanego wyznaczone na podstawie (1.33) oraz (1.34) są zbliżone (zob. Pintilie 2006; Klein, Moeschberger 2003; Kalbfleisch, Prentice 2002).

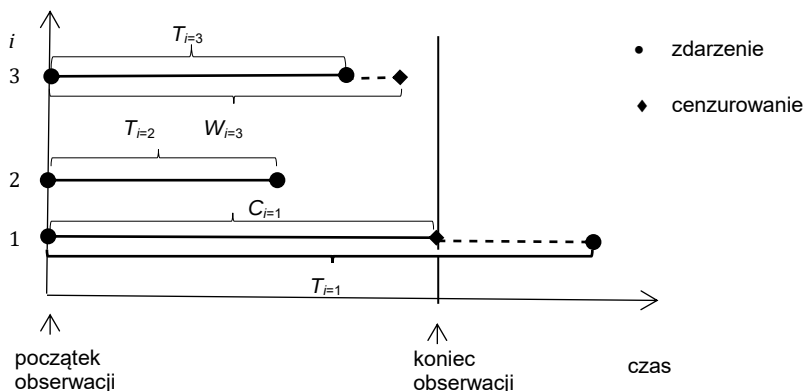
1.2. Cenzurowanie: rodzaje, modele cenzurowania

Analiza przetrwania ma zastosowanie w szczególności, gdy czasy trwania niewszystkich jednostek są obserwowalne. Ze względu na powód niemożności obserwacji czasów trwania wszystkich jednostek można mówić o **cenzurowaniu** (ang. *censoring*) lub **ucinaniu** (ang. *truncation*) danych. Cenzurowanie występuje wówczas, gdy wszystkie jednostki mogą dostać się do próby, a w przypadku niektórych jednostek czas obserwacji nie obejmuje w pełni ich czasu trwania. Ucinanie polega na braku możliwości badania niektórych jednostek, których czas trwania jest krótszy (ucina nie lewostronne) lub dłuższy (ucina nie prawostronne) niż ustalona wartość progowa (Jackowska 2013, s. 89). W niniejszej pracy rozpatrywany będzie tylko przypadek cenzurowania danych. Badanie, w którym wszystkie jednostki w próbie doznały zdarzenia, nazywane jest badaniem kompletnym (ang. *complete sample*), natomiast badanie, w którym czasy trwania dla części jednostek nie są obserwowalne, nazywane jest badaniem niekompletnym (ang. *incomplete sample*).

W zależności od tego, czy czas obserwacji jest nieznany z powodu zakończenia obserwacji przed wystąpieniem zdarzenia (końcowego), czy też obserwacja, że zaszło zdarzenie końcowe, jest przeprowadzana z opóźnieniem i nieznany jest rzeczywisty moment wystąpienia zdarzenia końcowego, rozróżnia się odpowiednio cenzurowanie prawostronne i lewostronne⁴. **Cenzurowaniem prawostronnym** określa

⁴ Trzecim typem cenzurowania jest cenzurowanie przedziałowe (ang. *interval censoring*), które występuje wtedy, gdy nie jest znany dokładny czas wystąpienia zdarzenia końcowego, ale znane są

się sytuację, gdy dla i -tej jednostki wiadomo jedynie, że $T_i > C_i$, gdzie C_i jest czasem cenzurowania. Natomiast w przypadku **cenzurowania lewostronnego**, $T_i < W_i$, gdzie W_i jest czasem lewostronnego cenzurowania jednostki i (Rossa 2005a, s. 14–16). Na wykresie 1.4 przedstawiono jednostki z czasem kompletnym ($i = 2$), cenzurowanym prawostronnie ($i = 1$) oraz cenzurowanym lewostronnie ($i = 3$). Dla jednostek cenzurowanych ($i = 1$ oraz $i = 3$) znane są tylko czasy cenzurowania, odpowiednio $C_{i=1}$ oraz $W_{i=3}$.



Wykres 1.4. Obserwacje kompletne i cenzurowane

Źródło: opracowanie własne.

Można wyróżnić cenzurowanie **typu I** oraz **typu II** (Cohen 1991, s. 4). W cenzurowaniu typu I czas cenzurowania jest z góry ustalony. W rezultacie w n -elementowej próbie liczby obserwacji kompletnych i cenzurowanych znane są dopiero w ustalonym momencie cenzurowania i należy je traktować jako realizacje zmiennej losowej. Natomiast w cenzurowaniu typu II ustalona jest z góry liczba obserwacji kompletnych, a przez to i liczba jednostek cenzurowanych. Obserwacja próby kończy się w momencie wystąpienia założonej liczby zdarzeń; czas, w którym to nastąpi, nie jest więc z góry znany (szerzej: Cohen 1991; Balicki 2006; Millard 2013). W zależności od tego, czy ustalony jest jeden moment cenzurowania, czy też jednostki są cenzurowane w jednym z wielu ustalonych, kolejnych momentów, wyróżnia się **cenzurowanie jednoetapowe** (ang. *single-stage censoring*) oraz **cenzurowanie stopniowe**, inaczej **progresywne** (ang. *multiple, progressive censoring*). W niniejszej pracy analizowany jest tylko przypadek cenzurowania prawostronnego i jednoetapowego.

Jeżeli w badaniu zaplanowany jest moment zakończenia badania, wynikające z tego cenzurowanie prawostronne nazywane jest **cenzurowaniem arbitralnym** (ang. *administrative censoring*) (Rossa 2005a). Jednostki mogą mieć równy lub różny

granice przedziału, w którym do zdarzenia doszło. Cenzurowanie przedziałowe nie będzie rozważane w niniejszej pracy.

czas cenzurowania arbitralnego w zależności od tego, czy dołączały do badania w tym samym, czy różnym czasie kalendarzowym. Jeżeli wszystkie jednostki przystępują do badania w tym samym momencie, czas cenzurowania $C_i = c$ jest jednokowy dla wszystkich jednostek i nie jest zmienną losową. Jeżeli jednostki przystępują do badania w różnych momentach i dodatkowo momenty przystępowania jednostek do badania są losowe, czas cenzurowania jest zmienną losową (szerzej: Balicki 2006; Jackowska 2013).

Cenzurowanie może wystąpić także z powodów niezależnych od badacza, m.in.:

- utraty jednostki z pola obserwacji (ang. *loss to follow-up*),
- wycofania się jednostki z badania (ang. *withdrawal*),
- wystąpienia zdarzeń konkurujących (ang. *competing events*).

W pierwszym przypadku obserwacja jednostki jest przerywana z powodów niezwiązanych z badanym zdarzeniem. Przykładowo w badaniach społecznych zmiana miejsca zamieszkania jest przyczyną utraty danej osoby z obserwacji. Utrata jednostek ma zwykle charakter losowy. Natomiast wycofanie się jednostki z badania to zwykle świadoma decyzja badanej jednostki lub badacza, powstająca często w związku z narażeniem na ryzyko wystąpienia badanego zdarzenia. Przykładowo w badaniach medycznych pacjent przestaje być częścią obserwowanej zbiorowości chorych, jeśli zostanie względem niego zmieniona metoda leczenia. Jeśli zmiana ta jest związana z pogorszeniem lub polepszeniem jego stanu zdrowia, ze względu na daną chorobę, wycofanie się z badania nie jest losowe. Zdarzenia konkurujące to zdarzenia, których wystąpienie uniemożliwia wystąpienie zdarzenia badanego. Zdarzenia konkurujące mogą być zarówno niezależne od zdarzenia badanego, jak i powiązane z rozkładem czasu trwania do wystąpienia zdarzenia badanego.

Większość metod analizy czasu trwania zakłada losowość oraz niezależność cenzurowania. **Cenzurowanie losowe** oznacza, że jednostki, które były cenzurowane w momencie t , są reprezentatywne pod względem rozkładu czasu trwania dla jednostek, które dotrwały do momentu t i nie są cenzurowane, tj. intensywność (hazard) dla jednostek cenzurowanych jest taka sama jak dla jednostek pozostających w próbie (Kleinbaum, Klein 2012). **Cenzurowanie niezależne** oznacza, że w dowolnie wybranej grupie jednostek w próbie te z jednostek, które są cenzurowane w momencie t , mają taki sam rozkład czasu trwania jak jednostki z danej grupy, które pozostają w próbie (Kleinbaum, Klein 2012). Cenzurowanie niezależne oznacza więc, że dla wyodrębnionej grupy jednostek, np. ze względu na pewną cechę inną niż czas trwania, cenzurowanie jest losowe w tej grupie. **Cenzurowanie nieinformacyjne/nieinformatywne** (ang. *non-informative censoring*) ma miejsce wówczas, gdy rozkład zmiennej T nie zawiera informacji o rozkładzie zmiennej C , i na odwrót (Kleinbaum, Klein 2012)⁵.

Cenzurowanie prowadzi do tego, że zmienna losowa T nie jest obserwowalna, a jej rozkład nie jest identyfikowalny. Odtworzenie tego rozkładu jest jednak możliwe, gdy przyjmie się model cenzurowania nieinformatywnego (Rossa 2005a, s. 19).

⁵ W pracy Rossa 2005a przedstawiono matematyczną definicję cenzurowania nieinformatywnego.

Jednym z modeli nieinformatywnych jest model cenzurowania losowego i niezależnego zdefiniowany przez Efrona (1967).

Niech $T_1, T_2, \dots, T_n$ będą czasami trwania o rozkładzie opisanym dystrybuantą F , funkcją trwania S oraz funkcją gęstości f . Z kolei niech $C_1, C_2, \dots, C_n$ będą niezależnymi, nieobserwowalnymi zmiennymi losowymi o identycznym rozkładzie opisanym dystrybuantą G , funkcją trwania $\bar{G}$ oraz funkcją gęstości g . Zmienne C_i są czasami cenzurowania związanymi z T_i (Balicki 2006, s. 76; Miller 1981, s. 4). W modelu losowym, niezależnym obserwowalna jest para zmiennych (Y_i, δ_i) , taka że czas obserwacji jest zmienną losową: $Y_i = \min(T_i, C_i)$ oraz

$$\delta_i = \mathbf{1}(T \leq Y) = \begin{cases} 1 & \text{gdy } T_i \leq C_i \\ 0 & \text{gdy } T_i > C_i \end{cases} \quad (1.35)$$

jest zmienną wskaźnikową informującą o tym, czy obserwacja jest kompletna, czy cenzurowana. Zmienne losowe Y_i są czasami obserwacji dla poszczególnych jednostek. Ze względu na to, że poszczególne jednostki mogą zaistnieć w próbie w sposób losowy, jak i w sposób losowy może być zakończona ich obserwacja, zmienne Y_i są traktowane jako zmienne niezależne o identycznym rozkładzie⁶. Stanowi to konsekwencję tego, że zmienne T_i oraz C_i są niezależne.

1.3. Funkcja wiarygodności dla danych cenzurowanych

Funkcja wiarygodności to prawdopodobieństwo otrzymania zaobserwowanej próby. Postać funkcji wiarygodności zależy od tego, czy chodzi o badanie kompletne, czy cenzurowane. Jako pierwsza zostanie przedstawiona podstawowa postać funkcji wiarygodności, tj. postać dla badań kompletnych (Fisz 1969).

Niech $F(x)$ będzie dystrybuantą zmiennej losowej X . Niech dystrybuanta ta zależy od m nieznanych parametrów α_k ($k = 1, 2, \dots, m$), które mają być oszacowane na podstawie próby n elementowej $(x_1, \dots, x_n)$. Dla zmiennej losowej X , która jest typu skokowego z prawdopodobieństwami realizacji $P(X = x_i) = f(x_i)$, gdzie $(i = 1, 2, \dots, n)$, które są funkcjami parametrów α_k , oraz dla zmiennej X typu ciągłego o gęstości $f(x)$, gdzie $f(x_i)$ są funkcjami parametrów α_k , funkcja wiarygodności (wiarogodności) próby jest dana wzorem

$$L = \prod_{i=1}^n f(x_i). \quad (1.36)$$

W badaniach cenzurowanych funkcja wiarygodności musi być zmodyfikowana ze względu na występowanie obserwacji cenzurowanych i niesioną przez nie niepełną

⁶ Zmienne losowe X i Y o rozkładach określonych przez dystrybuanty, odpowiednio $F_1(x)$ oraz $F_2(y)$, są niezależne, jeśli dla dowolnej pary (x, y) zachodzi równość: $F(x, y) = F_1(x)F_2(y)$, gdzie $F(x, y)$ jest dystrybuantą zmiennej dwuwymiarowej (X, Y) . Gdy zmienne są niezależne, rozkład warunkowy jednej zmiennej jest jednakowy dla wszystkich wartości drugiej zmiennej. Innymi słowy, jakkolwiek uzyskana wartość jednej zmiennej nie daje żadnych informacji o rozkładzie drugiej zmiennej (Fisz 1969).



Wydawnictwo
Uniwersytetu Gdańskiego

ISBN 978-83-7865-826-9