Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka stosowana
Specjalność: –

Praca dyplomowa – inżynierska

TESTOWANIE ZGODNOŚCI Z ROZKŁADAMI O LEKKICH I CIĘŻKICH OGONACH ZA POMOCĄ FUNKCJI NADWYŻKI SZKODY

Piotr Mikler

słowa kluczowe:
funkcja nadwyżki szkody, testowanie, ogon rozkładu, rozkłady ciężkoogonowe, rozkłady lekkoogonowe

krótkie streszczenie:
W niniejszej pracy zaproponowany jest test statystyczny klasyfikujący ogon próbki jako cięższy, lub lżejszy od ogonu rodziny rozkładów wykładniczych. Badana jest funkcja nadwyżki szkody dla próby losowej z rozkładu wykładniczego. Wyprowadzamy rozkład tej statystyki na którego podstawie konstruujemy test statystyczny. Proponujemy kryterium wyboru parametru maksymalizującego jego moc. Prezentujemy działanie testu w znanych klasach rozkładów oraz analizujemy z jego pomocą szkodowe dane ubezpieczeniowe.

Opiekun pracy dyplomowej	dr hab. inż. Krzysztof Burnecki	.....	.....
	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 2021

Wrocław University
of Science and Technology

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

Engineering Thesis

TESTING OF LIGHT AND HEAVY-TAILED DISTRIBUTIONS WITH MEAN EXCESS LOSS FUNCTION

Piotr Mikler

keywords:
mean excess function, testing, distribution tail, heavy-tailed distribution, light-tailed distribution

short summary:
In the following thesis we propose a statistical test classifying a distribution's tail as heavier or lighter than tails of exponential distributions. Considering a sample drawn from an exponential distribution we derive it's empirical mean excess function distribution. We use this statistic to construct a statistical test, and we propose an algorithm for choosing parameters that maximize the test's power. We illustrate the performance of the test for chosen distribution classes and apply the test to insurance loss data.

Supervisor	dr hab. inż. Krzysztof Burnecki	.....	.....
	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, 2021

Spis treści

Wstęp

Każdego dnia każdy z nas spotyka się z milionami losowych zdarzeń i procesów. Determinują one otaczający nas świat, nasze samopoczucie, wydarzenia bliskie i odległe w mniejszym lub większym stopniu. Najczęściej są one tak subtelne, że umyka naszej uwadze ich losowa natura. Nikt przecież nie zastanawia się co spowodowało, że za oknem mamy dziś temperaturę dwunastu stopni Celsjusza, albo dlaczego w korku przed nami stoi akurat pięć samochodów. Na codzień mijamy te zdarzenia obojętnie, nie zastanawiając się nad ich naturą, a okazujemy zainteresowanie dopiero gdy przyjmują ekstremalną formę. Duże, medialne wydarzenia jak krach giełdowy, rekordowe opady i powodzie, kilometrowe korki - to przecież nic innego jak wzmocnione odmiany codziennych spadków cen akcji, przelotnych deszczów czy kilku samochodów na czerwonym świetle.

Czy każde losowe zdarzenie ma szansę tak okazale zaprezentować swoją moc? Niektóre zjawiska wydają się przecież mieć pewne naturalne ograniczenie. Opisując zjawiska losowe można dojść do wniosku, że istnieją w takim razie co najmniej dwa rodzaje rozkładów zmiennych losowych. Takie, w których powinniśmy się spodziewać odstających obserwacji z istotnym prawdopodobieństwem, oraz takie gdzie pojedyncze obserwacje mogą odchylać się od reszty, lecz są to rzadkie i nieprawdopodobne sytuacje. Czy da się odróżnić od siebie te dwa rodzaje zmiennych losowych na podstawie pojedynczej próbki?

W niniejszej pracy staramy się znaleźć odpowiedź na to pytanie z zastosowaniem metod statystycznych. W pierwszym rozdziale skupiamy się na matematycznym modelu losowych wartości jakim jest zmienna losowa. Opisujemy sposób jej definiowania i jej podstawowe charakterystyki. Następnie przyglądamy się pojęciu ogonu rozkładu - charakterystyce która opisuje częstość występowania ekstremalnych wartości. Wskazujemy na definicję ogonu, która wykorzystuje rozkład wykładniczy jako klasę oddzielającą rozkłady ciężkoogonowe od lekkoogonowych. W drugim rozdziale prezentujemy funkcję nadwyżki szkody oraz jej empiryczny estymator. Pomoże nam on badać relatywną ciężkość ogonu w odniesieniu do ogonów rozkładów wykładniczych, co pozwoli na klasyfikację rozkładu do jednej z wyżej wymienionych grup. W kolejnych dwóch rozdziałach można przeczytać o wybranych rodzinach zmiennych losowych charakteryzujących się różnymi ciężkościami ogonów. Prezentujemy sposoby na ich symulowanie, oraz estymację parametrów. W rozdziale piątym wyprowadzamy rozkład estymatora funkcji nadwyżki szkody dla przypadku rozkładu wykładniczego. Ten wynik pozwala skonstruować test statystyczny na rozkład wykładniczy, który przy odrzuceniu hipotezy zerowej klasyfikuje ogon rozkładu. Rozdział szósty ukazuje działanie testu dla znanych klas rozkładów, oraz dla danych rzeczywistych opisujących wypłaty ubezpieczyciela z tytułu ubezpieczeń odpowiedzialności cywilnej oraz autocasco.

Rozdział 1

Rozkład i jego ogon

W 1960 roku amerykański matematyk i fizyk Edward Lorenz pracował nad teorią, która na zawsze zapisała go na kartach historii. Opracowując prototyp komputerowego programu użył prostych równań opisujących zależności między meteorologicznymi wielkościami takimi jak ciśnienie atmosferyczne, temperatura, wilgotność i inne w celach stworzenia pionierskiej, wspomaganej mocą obliczeniową komputera prognozy pogody. Niestety gdy projekt był gotowy, ku jego rozczarowaniu okazało się, że dzieło nie ma najlepszej mocy predykcyjnej. Przeciwnie - mimo, że konstrukcja modelu była w mniemaniu Lorenza poprawna, to zaledwie delikatna zmiana parametrów wejściowych powodowała otrzymanie skrajnie różnych wyników [13].

Lorenz nie wiedział jeszcze wtedy, że mimo fiaska na polu meteorologii ta sama praca będzie punktem zapalnym badań nad nową gałęzią nauk matematycznych. Zaledwie kilka lat później uczony stał się bowiem ojcem teorii chaosu. Teorii według której nawet deterministyczne, prognozowalne układy ulegają tak zwanemu efektowi motyla - czyli na długim okresie czasu pod wpływem wielu bodźców zachowują się praktycznie losowo. W praktyce oznacza to, że jeśli mamy układ i delikatnie zmienimy jego warunki początkowe, to te pozornie małe różnice będą się z czasem akumulować, nierzadko rekurencyjnie wpływając na system. W związku z tym, dokładne przewidywanie często nawet prostych układów i systemów staje się nie tyle skomplikowane co niemożliwe. Jednym z przykładów niech będzie ruch wahadła podwójnego [9] - pozornie prostego w konstrukcji połączenia dwóch wahadeł fizycznych, które wynikowo zachowują się praktycznie nieprzewidywalnie.

Jeżeli problem potrafi sprawić nawet wahadło podwójne - to co dopiero można powiedzieć np. o globalnym systemie ekonomicznym? O grze siedmiu miliardów konsumentów, wielu rządów i organizacji. Jak opisać wynik działania czegoś wystawionego codziennie na tak niewyobrażalną liczbę bodźców? Nawet jeżeli jest on prognozowalny, to ciężko jest bronić tezy, że nie jest chaotyczny. Analizując skomplikowane systemy, często musimy zatem porzucić deterministyczne myślenie, na rzecz probabilistyki.

1.1 Zmienna losowa

Matematycznym modelem i podstawowym narzędziem które służy do opisu niedeterministycznych wielkości jest zmienna losowa.

Definicja 1.1 (Zmienna losowa)**.** [12] Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną, czyli niech Ω to pewien niepusty zbiór, $\mathcal{F}$ to σ-ciało zdarzeń losowych, a P to funkcja $P\colon \Omega \to [0,1]$. Zmienną losową X nazywamy wtedy funkcję $X\colon \Omega \to \mathbb{R}$ taką, że dla

dowolnego zbioru borelowskiego $\mathcal{B} \subset \mathbb{R}$ mamy:

$$\{\omega \colon X(\omega) \in \mathcal{B}\} \in \mathcal{F}.$$

Aby prawidłowo opisać zmienną losową potrzebujemy jeszcze funkcji, która jednoznacznie zdefiniuje jej rozkład.

Definicja 1.2 (Dystrybuanta)**.** [12] Dystrybuantą zmiennej losowej X nazywamy funkcję:

$$F_X(x) = P(X < x).$$

Dystrybuanta spełnia ten warunek opisując prawdopodobieństwo, że X będzie mniejszy niż jej argument. Niestety nie zawsze ma ona wygodną postać. Często definiowana jest całką, lub jest złożeniem pewnych funkcji specjalnych. Wtedy przydatne są inne sposoby opisu, jak np. gęstość, czy funkcja charakterystyczna rozkładu.

Nie zawsze też interesuje nas kompletny opis zmiennej losowej. Niejednokrotnie w praktyce potrzebujemy jedynie ogólnych miar jak np. średnia, czy odchylenie standardowe.

Definicja 1.3 (Wartość oczekiwana)**.** [12] Wartość oczekiwana ciągłej zmiennej losowej X definiowana jest jako:

$$\mathbb{E}[X] = \int_\Omega X(\omega) dP(\omega) = \int_{\mathbb{R}} x f_X(x) dx,$$

gdzie $f_X(x) = \frac{dF_X}{dx}(x)$ to gęstość zmiennej losowej X.

Bardziej ogólnie, jeśli weźmiemy $\mathcal{F}$-mierzalną funkcję $g(x)$ (tzn. $g(X) \in \mathcal{F}\ \forall_{X \in \mathcal{F}}$), to można napisać:

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) dx. \tag{1.1}$$

Wartość oczekiwana jest bardzo istotną informacją o rozkładzie, ponieważ jeśli istnieje, to właśnie do niej dąży średnia próbkowa. Nie dla każdego rozkładu niestety możemy prawidłowo ją zdefiniować. Gęstość rozkładu $f_X(x)$ musi zanikać wystarczająco szybko, żeby całka w definicji 1.3 była zbieżna. Jako przykład wystarczy wziąć chociażby klasę rozkładów stabilnych, w których odpowiednie wartości parametrów powodują istnienie, lub brak tej statystyki [1].

Średnia rozkładu jest jednak zaledwie reprezentantem dużo większej grupy charakterystyk jaką są momenty rozkładu.

Definicja 1.4 (Momenty rozkładu)**.** [12] Moment rzędu p zmiennej losowej X definiujemy jako wielkość

$$\mathbb{E}[X^p] = \int_\Omega X^p dP.$$

Momentem centralnym rzędu p nazywamy natomiast

$$\mathbb{E}[X^p] = \int_\Omega (X - \mathbb{E}X)^p dP.$$

W szczególności moment centralny drugiego rzędu nazywamy wariancją i oznaczamy jako

$$\mathbb{V}ar[X] = \int_\Omega (X - \mathbb{E}[X])^2 dP = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Można dodatkowo udowodnić, że jeżeli istnieje moment rzędu p, to istnieją również wszystkie momenty mniejszych rzędów. Z tego powodu jeśli dla rozkładu nie istnieje średnia, to tym bardziej nie istnieje wariancja i cała reszta momentów o wyższych rzędach.

Alternatywnie momenty rozkładu można jednoznacznie opisać stosując funkcję tworzącą (generującą) momenty. Jest to funkcja wyrażona poprzez wartość oczekiwaną, a jej główną zaletą jest specjalna postać rozwinięcia Taylora.

Definicja 1.5 (Funkcja tworząca momenty)**.** [2] Dla zmiennej losowej X funkcją tworzącą momenty nazywamy funkcję:

$$M_X(t) = \mathbb{E}[e^{Xt}].$$

Lemat 1.6. *Jeżeli istnieje funkcja generującą momenty $M_X(t)$ zmiennej losowej X, to moment rzędu p można obliczyć korzystając z równości:*

$$\mathbb{E}[X^p] = M_X^{(p)}(0),$$

pod warunkiem, że $M_X^{(p)}(0) < \infty$.

Dowód. Rozwijając funkcję $M_X(t)$ w szereg Maclaurina otrzymujemy:

$$M_X(t) = 1 + \mathbb{E}[X]t + \frac{\mathbb{E}[X^2]t^2}{2!} + \frac{\mathbb{E}[X^3]t^3}{3!} + \dots$$

Zatem n-ta pochodna tej funkcji da się przedstawić jako:

$$M_X^{(n)}(t) = \mathbb{E}[X^n] + \frac{\mathbb{E}[X^{n+1}]t}{1!} + \frac{\mathbb{E}[X^{n+2}]t^2}{2!} + \dots$$

Dla $t = 0$ otrzymujemy zatem równość:

$$M_X^{(n)}(0) = \mathbb{E}[X^n].$$

□

Momenty rozkładu są istotne z praktycznego punktu widzenia. Nie dają kompletnej informacji o nim, lecz znajomość momentów wyższych rzędów dość rygorystycznie opisuje zachowanie ogonu, co może być przy niektórych problemach wystarczające.

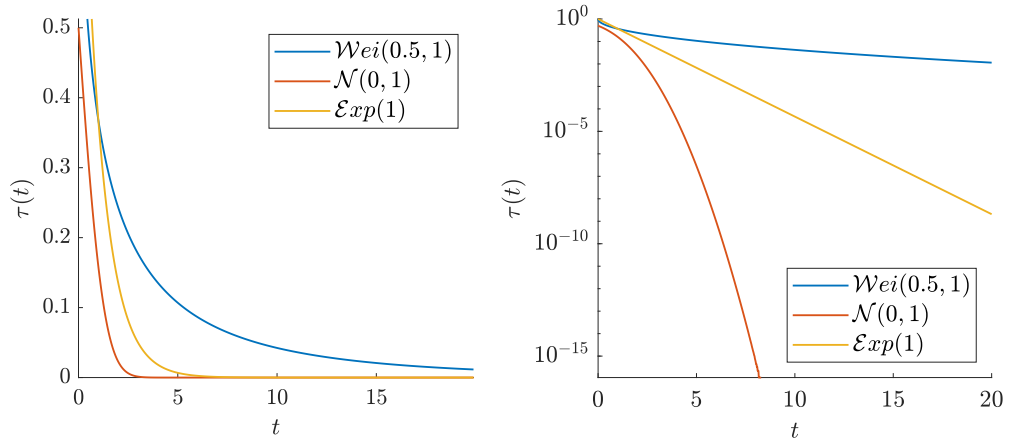
Praca z rozkładami zmiennych losowych pozwala obserwować całą paletę ich rozmaitych, często nietrywialnych zachowań i własności (od kurtozy i skośności, przez rozkład maksimum, aż po brak pamięci). Chciałbym zwrócić w tym miejscu uwagę czytelnika na fascynujący według mnie fakt, jakim jest to, że wszystkie te własności wynikają tylko i wyłącznie ze sposobu opisu pary wartość-prawdopodobieństwo.

1.2 Ogon rozkładu

Charakterystyką rozkładu która jest tematem przewodnim tej pracy jest ciężkość ogonu rozkładu. Ogon opisuje zachowanie zmiennej losowej dla wartości odległych od centrum rozkładu. Graficzną reprezentacją tej cechy jest funkcja ogonu.

Definicja 1.7 (Funkcja ogonu rozkładu)**.** [1] Funkcją ogonu rozkładu zmiennej losowej X nazywamy

$$\tau(x) = P(X > x) = 1 - F_X(x).$$



Rysunek 1.1: Ogony rozkładów: ciężkoogonowego $\mathcal{W}ei(0.5,1)$, $\mathcal{N}(0,1)$ oraz $\mathcal{E}xp(1)$. Lewy panel - funkcje ogonu tych rozkładów. Prawy panel - funkcje ogonu w skali logarytmicznej

Ta funkcja pozwala bardzo intuicyjnie spojrzeć na ogon rozkładu oraz łatwo porównywać grubości różnych ogonów. Im grubszy ogon ma rozkład, tym większe wartości przyjmuje funkcja. Najczęściej w rozważanych modelach ciągłych obserwujemy monotoniczne, stopniowe jej zanikanie. Analiza ciężkości ogonu skupia się na tempie tego zaniku. Jeśli jest ono odpowiednio duże, to funkcja szybko gaśnie sprawiając, że duże realizacje zmiennej losowej będą bardzo rzadkie, a rozkład w konsekwencji będzie lekkoogonowy. Dla rozkładów ciężkoogonowych obserwujemy natomiast odwrotną sytuację - funkcja ogonu zanika powoli, powodując istotne szanse na duże realizacje.

W pracy rozważać będziemy bez straty ogólności jedynie prawy ogon rozkładu, lecz te same rozważania możnaby przeprowadzić dla lewego ogonu dokonując transformacji zmiennej losowej z X do Y poprzez $Y := -X$.

Pojęcie „powoli", lub „szybko" zanikającego ogonu jest intuicyjne i zrozumiałe, jednak regułę klasyfikującą ogony można matematycznie ująć na wiele sposobów. Z tego powodu obecnie brakuje jednogłośnie zaakceptowanej definicji ciężkich ogonów, a w użyciu najczęściej pojawiają się trzy z nich.

Pierwsza i najprostsza definicja wykorzystuje pojęcie leptokurtycznych rozkładów i porównuje ogon rozkładu do rozkładu normalnego.

Definicja 1.8 (Kurtoza)**.** [4] Kurtozą rozkładu nazywamy wielkość:

$$\kappa_X = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathbb{V}ar[X]^2} - 3.$$

Dla rozkładu normalnego tak zdefiniowana wielkość jest zerem[1]. Rozkłady o kurtozie większej od zera nazywamy leptokurtycznymi i rzeczywiście charakteryzują się ogonami cięższymi od rozkładu normalnego. Jest to najmniej wymagająca z definicji ciężkiego ogonu, w związku z czym w tej grupie znajdziemy wiele rozkładów, np. rozkład t-studenta, rodzinę rozkładów wykładniczych, czy rozkłady Laplace'a.

Druga z definicji [4] odwołuje się bezpośrednio do funkcji ogonu. Punktem wyjściowym do jej sformułowania jest następująca własność ogonu:

$$\tau(x) \sim x^{-\alpha}, \text{ dla } x \to \infty, \ \alpha > 0. \tag{1.2}$$

[1]Czasem w literaturze definicja nie jest odpowiednio korygowana, co powoduje wartość 3.

Definicja postuluje, że jeśli ogon ma tę własność (*power law*), to rozkład możemy nazywać ciężkoogonowym. Łatwo porównywać grubość tak zdefiniowanych ciężkich ogonów, bo im większy jest parametr α tym szybciej zanikający (lżejszy) jest ogon.

Lemat 1.9 (Istnienie momentów a ogon rozkładu)**.** *Dla rozkładu o ogonach zanikających asymptotycznie jak $x^{-\alpha}$ nie istnieją momenty rzędów $p \geqslant \alpha$.*

Dowód. Dla pewnej stałej $A \in \mathbb{R}$, oraz wystarczająco dużej stałej $M \in \mathbb{R}$ rozpatrzmy całkę niewłaściwą postaci:

$$A \int_M^\infty \frac{1}{x^\alpha} dx.$$

Wiemy o niej, że jest zbieżna wtedy i tylko wtedy, gdy $\alpha > 1$. Załóżmy ponadto, że ogon rozkładu ma własność:

$$1 - F_X(x) \sim x^{-\alpha}, \text{ dla } x \to \infty.$$

To implikuje z kolei

$$x^p f_X(x) \sim \alpha x^{-\alpha+p-1}.$$

Widzimy zatem, że rozpatrując całkę w poszukiwaniu p-tego momentu spotykamy się z wyrażeniem, które dla odpowiednio dużego M zachowuje się jak

$$\int_M^\infty x^p f_X(x) dx \sim \alpha \int_M^\infty x^{-\alpha+p-1} dx.$$

Całka po prawej stronie będzie zbieżna jedynie gdy $p < \alpha$, co oznacza istnienie momentów wyłącznie o rzędach mniejszych niż α. □

Kontrastując tę definicję z poprzednią, zauważmy, że jeżeli ogon spełnia równanie (1.2) dla $\alpha < 1$, to rozkład nie posiada wartości oczekiwanej, więc na mocy lematu 1.9 nie może mieć ani wariancji, ani czwartego momentu centralnego. Kurtoza nie będzie więc wtedy poprawnie zdefiniowana. Wynika z tego, że definicja wykorzystująca *power law* może być stosowana do szerszej grupy rozkładów.

Mamy w końcu również trzecią definicję która określa jako rozkłady ciężkoogonowe rozkłady które nie są ograniczone funkcją wykładniczą.

Definicja 1.10 (Rozkład ciężkoogonowy)**.** [3] Jeżeli ogony rozkładu zmiennej losowej X zanikają wolniej niż ogon rozkładu wykładniczego, tzn.

$$\forall_{\lambda>0} \quad \lim_{x\to\infty} \frac{\tau_X(x)}{e^{-\lambda x}} = \infty,$$

to rozkład X nazywamy ciężkoogonowym. Alternatywnie, rozkład nazywamy ciężkoogonowym, gdy nie istnieje dla niego funkcja generująca momenty, czyli

$$\forall_{t>0} \quad M_X(t) = \infty.$$

Jeżeli przypomnimy sobie, że dystrybuanta rozkładu wykładniczego o parametrze λ dana jest wzorem:

$$F_X(x) = 1 - e^{-\lambda x},$$

to łatwo zauważyć, że w mianowniku definicji 1.10 znajduje się funkcja ogonu tego rozkładu. W zaprezentowanej wyżej definicji zatem to rozkład wykładniczy pełni rolę granicy. Ogony

zanikające wolniej od wykładniczych powodują że iloraz w definicji 1.10 rozbiega do nieskończoności, a rozkład X przez to trafia do grupy ciężkoogonowych. Rozkłady o szybszym zaniku powodują zanik ilorazu do zera i uznawane są za lekkoogonowe.

Atrakcyjną cechą takiej definicji ciężkości ogonu jest fakt, że gdy przedstawimy go na wykresie przy skali logarytmicznej to możemy obserwować transformację ogonu wykładniczego na funkcję liniową $y = 1 - \lambda x$ (widoczne na rysunku 1.1). Po takim przekształceniu można porównywać ciężkość ogonu z ogonem wykładniczym gołym okiem. Dodatkowo zauważmy, że ta definicja nie wymaga od rozkładu posiadania żadnego z momentów, potrzebujemy jedynie znać asymptotyczne zachowanie dystrybuanty rozkładu X.

Jak widzimy nie istnieje jedna słuszna definicja ciężkości ogonu. Każda z wyżej wymienionych ma swoje racje i uchwyca w swoim własnym sensie reżym ogonów ciężkich i lekkich. Nie powinno ulegać jednak wątpliwości, że takie dwie grupy rozkładów istnieją i należy liczyć się z grubością ogonu rozkładu, szczególnie w zastosowaniach praktycznych. Złe dopasowanie może bowiem prowadzić długoterminowo do niechcianych rezultatów, błędów predykcji czy nawet materialnych strat [7]. My na potrzeby tej pracy skupimy się na definicji 1.10. Pisząc o ogonie ciężkim lub lekkim będziemy mieć zatem na myśli odpowiednio ogon cięższy lub lżejszy od ogonu rozkładu wykładniczego.

Rozdział 2

Funkcja nadwyżki szkody

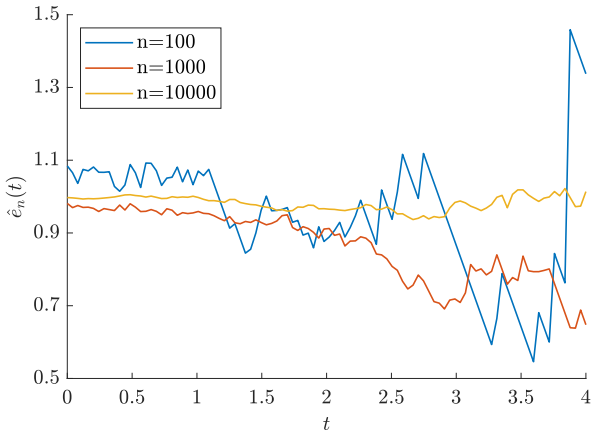
Jako narzędzie do badania ogonu rozkładu posłuży nam funkcja nadwyżki szkody (z ang. *Mean Excess Function*, MEF). Określa ona średnią rozkładu ponad pewien poziom odcięcia, pod warunkiem jego przekroczenia.

Definicja 2.1 (Funkcja nadwyżki szkody (MEF))**.** [2] Funkcją nadwyżki szkody zmiennej losowej X nazywamy

$$e_X(t) = E[X - t | X > t].$$

Przykładowe trajektorie tej funkcji obejrzeć można na rysunku 4.1. Funkcja nadwyżki szkody jest powszechnie znana przez praktyków w wielu obszarach nauki. W zależności od kontekstu w którym jest używana, funkcja ta ma wiele naturalnych interpretacji i nierzadko w specjalistycznych środowiskach spotyka się ją pod innymi nazwami. W dziedzinie zarządzania ryzykiem finansowym aby badać ekspozycję firmy na możliwe potknięcia finansowe bada się rozkłady zmiennych losowych typu „Cashflow", „Profit/Loss", czy po prostu „Loss". Jedną z miar popularnych wśród praktyków służącą do tych analiz jest *Expected Shortfall (ES)* [7]. Funkcja ta mierzy średnią stratę firmy zakładając przebicie się jej ponad pewien poziom. Tak zdefiniowana *ES* jest jednak niczym innym jak funkcją nadwyżki szkody skorygowaną o liniowy trend.

Jeżeli natomiast weszlibyśmy w teorię ubezpieczeń życiowych a przez X oznaczylibyśmy czas życia człowieka, to przy problemach związanych z wyceną polis, rent czy świadczeń



Rysunek 2.1: Estymator funkcji nadwyżki szkody dla rozkładu $\mathcal{E}xp(1)$. Porównanie zbieżności dla różnych długości próbek n.

będziemy używać funkcji średniego pozostałego czasu życia *„Mean Residual Life" (MRL)* [6]. Niesie ona informację o średnim pozostałym czasie życia osoby w wieku t. Taka informacja jest jednym z kluczy pozwalających teoretycznie na „sprawiedliwą" wycenę składki ubezpieczeniowej. Łatwo zauważyć, że średni pozostały czas życia to intuicyjna interpretacja funkcji nadwyżki szkody zaaplikowanej do zmiennej losowej oznaczającej długość życia człowieka. Jest to zaledwie para przykładów, lecz pokazują one jak bardzo interdyscyplinarna jest koncepcja funkcji nadwyżki szkody.

Choć analityczna funkcja *MEF* jest dobrze zdefiniowana i intuicyjna dla rozkładów ciągłych, w praktycznych zastosowaniach zmuszeni jesteśmy do używania estymatora $\hat{e}_n(t)$ wyliczanego na podstawie próbki $[x_1, x_2, \ldots, x_n]$:

Definicja 2.2 (Estymator funkcji nadwyżki szkody)**.** [2] Dla próbki losowej $[x_1, x_2, \ldots, x_n]$ estymatorem funkcji nadwyżki szkody nazywamy funkcję:

$$\hat{e}_n(t) = \frac{\sum_{\{x_i > t\}} x_i}{\#\{i : x_i > t\}} - t.$$

Warto zwrócić uwagę na fakt, że estymator zaprezentowany w definicji 2.2 zdefiniowany jest tylko dla $t \leqslant \max_i\{x_i\}$. W kolejnych rozdziałach będziemy badać rozkład wartości estymatora i wygodne będzie zdefiniowanie go również na obszarze $(\max_i\{x_i\}; \infty)$. Biorąc pod uwagę, że

$$\lim_{t \downarrow \max_i\{x_i\}} \hat{e}_n(t) = 0,$$

oraz interpretację *MEF* jako średnią nadwyżkę powyżej t zdecydowaliśmy przypisać tej funkcji wartość 0 w miejscach w których nie jest zdefiniowana. Ograniczając się bowiem tylko i wyłącznie do informacji oferowanych nam przez próbkę maksymalną wartością jaką możemy dostać jest $\max_i\{x_i\}$, więc średnia nadwyżka ponad tę wartość (według całej dostępnej nam informacji) powinna wynieść zero.

Jako efekt otrzymujemy zamiast estymatora przedstawionego w definicji 2.2 nową definicję, określoną na całej przestrzeni liczb rzeczywistych.

Definicja 2.3 (Estymator funkcji nadwyżki szkody)**.** Jeżeli $[x_1, x_2, \ldots, x_n]$ to wektor realizacji pewnej zmiennej losowej, to estymatorem funkcji nadwyżki szkody nazywamy dla niego funkcję:

$$\hat{e}_n(t) = \left(\frac{\sum_{x_i > t} x_i}{\#\{i : x_i > t\}} - t \right) \mathbb{1}_{\{t < \max_i\{x_i\}\}}.$$

Empiryczny estymator takiej postaci zbiega do swojego analitycznego odpowiednika prawie na pewno wraz z $n \to \infty$ [5]. Jego trajektorie to odcinki, o skokach w miejscach zadanych próbką, oraz liniowym zanikiem pomiędzy nimi (rysunek 2.1). Ponadto zauważmy, że dla ustalonego t estymator jest średnią pewnego podzbioru próbki, pomniejszoną o ten parametr. O parametrze t można myśleć jako o punkcie odcięcia, ponieważ tylko obserwacje większe od niego są używane przy obliczaniu estymatora. Im większa jego wartość, tym mniej próbek spełnia warunek, a co za tym idzie estymator jest liczony z mniejszej liczby obserwacji. Powoduje to, że wariancja wartości $\hat{e}_n(t)$ nie może maleć wraz ze wzrostem t. Widać to dobrze rysunku 2.1, gdzie obserwujemy coraz wyraźniejsze wahania estymatora dla coraz większych t. Wielkość tych wahań istotnie zależy od długości próbki. Im więcej elementów ma próbka, tym więcej obserwacji jest brana pod uwagę, co powoduje większą stabilność estymatora i lepszą zbieżność do wartości teoretycznej.

Poza wyżej wymienionymi własnościami funkcja nadwyżki szkody jest również sposobem opisu dynamiki zaniku ogonu. Zauważmy, że wraz ze wzrostem ciężkości ogonu wzrasta domieszka dużych wartości w rozkładzie. To z kolei powoduje, że średnia w definicji 2.2 będzie większa, co pociąga za sobą również większą wartość funkcji *MEF*. Również w drugą stronę - szybki zanik ogonu będzie powodował mniejsze wartości funkcję nadwyżki szkody. W kolejnym rozdziale pokażemy na dodatek, że funkcja *MEF* w przypadku rokładu wykładniczego jest funkcją stałą. Intuicyjnie widać zatem, że funkcja nadwyżki szkody może być narzędziem detekcji ciężkości ogonów zgodnie z definicją 1.10.

Rozdział 3

Estymatory parametrów rozkładu i podstawy symulacji

3.1 Metody symulacji zmiennych losowych

Współczesna matematyka, a co za tym idzie również ta praca w znaczącym stopniu opiera się na komputerowych symulacjach. W poniższym rozdziale przedstawimy dwa podstawowe sposoby symulowania zmiennych losowych o zadanym rozkładzie: metodę odwrotnej dystrybuanty oraz metodę akceptacji-odrzucenia.

Algorytm 3.1 (Metoda odwrotnej dystrybuanty)**.** *[11] Załóżmy, że jesteśmy w stanie symulować zmienną losową z rozkładu jednostajnego* $U \sim \mathcal{U}(0;1)$*.*

Cel: *Symulacja zmiennej zgodnie z dystrybuantą* $F_X(x)$*.*

1. *Znajdź funkcję* $F_X^{-1}(x)$*.*
2. *Wylosuj* $u \sim \mathcal{U}(0;1)$*.*
3. $Y \coloneqq F^{-1}(u)$*.*

Tak tworzona zmienna Y *ma rozkład zadany dystrybuantą* $F_X(x)$*.*

Ta metoda sprawdza się dobrze jeżeli możemy łatwo odwrócić dystrybuantę. Jest również najwygodniejszą z metod symulacji, ponieważ cały proces sprowadza się do wysymulowania tylko jednej zmiennej jednostajnej i przetransformowania jej pewną deterministyczną funkcją. Często jednak spotykamy się z sytuacjami w których mamy dostępną nie dystrybuantę, lecz gęstość rozkładu. Możemy wtedy zaimplementować inną metodę symulacji zmiennej losowej.

Algorytm 3.2 (Metoda akceptacji-odrzucenia)**.** *[11] Załóżmy, że jesteśmy w stanie symulować: zmienną losową o rozkładzie jednostajnym* $U \sim \mathcal{U}(0;1)$*, zmienną losową* Y *z rozkładu o gęstości* $g_Y(x)$*, oraz, że istnieje stała* $c \geqslant \sup_x\{g_Y(x)/f_X(x)\} \geqslant 1$*.*[1]

Cel: *Symulacja zmiennej losowej o funkcji gęstości* $f_X(x)$*.*

1. *Wylosuj* $y \sim g_Y(x)$*.*
2. *Wylosuj* $u \sim \mathcal{U}(0;1)$*:* $\quad u \perp y$*.*

[1] Algorytm będzie działał najszybciej, jeśli $c \approx 1$.

3. *Jeżeli* $f_X(y) \geqslant c \cdot u \cdot g_Y(y)$, *to wstaw* $X \coloneqq y$.

   W przeciwnym wypadku wróć do kroku 1.

Wygenerowana w ten sposób zmienna X *ma zadany rozkład* $f_X(x)$.

Przedstawiony powyżej algorytm pozwala wygenerować zdecydowaną większość rodzin rozkładów spotykanych w praktyce. Z tego samego powodu okazuje się jednak często nieoptymalny. Dlatego zawsze warto sprawdzić czy nie mamy możliwości użycia innego algorytmu, dedykowanego dla żądanego rozkładu.

Nierzadko zdarza się również, że nie chcemy zakładać nic o rozkładzie badanej zmiennej, lecz losować realizacje zgodnie z empirycznym rozkładem pewnej dostępnej próbki. Dokonuje się tego najprościej przez metodę odwrotnej dystrybuanty, przy drobnej modyfikacji polegającej na wprowadzeniu w miejsce analitycznej dystrybuanty jej empirycznego estymatora (często dodatkowo wygładzonego). To podejście jest popularne np. w zarządzaniu ryzykiem [7], gdy zależy nam na uniknięciu błędu związanego z wyborem modelu i wolimy generować realizacje/scenariusze na podstawie historycznych wartości.

3.2 Estymatory i metody estymacji

Przy wnioskowaniu statystycznym nie unikniemy problemu doboru parametrów. Mając dostępny wektor danych $[x_1, \ldots, x_n]$ należy dobrać je do modelu w taki sposób, aby jak najlepiej dopasował się on do danych. Jedną z najbardziej znanych metod estymacji jest metoda największej wiarogodności (*Maximum Likelihood Estimation, MLE*). Estymatory otrzymane tą metodą *MLE* mają szereg atrakcyjnych własności (efektywność, zgodność, asymptotyczna nieobciążoność) [8] które sprawiają, że sięga się po nie gdy tylko są dostępne.

Algorytm 3.3 (Metoda największej wiarogodności)**.** *[8] W celu otrzymania estymatora* $\hat{\theta}$ *pewnego parametru* θ *metodą MLE należy zacząć od znalezienia funkcji wiarogodności, czyli funkcji gęstości łącznej* $f_n(x_1, \ldots, x_n)$ *rozkładu liczonej w punktach zadanych przez otrzymaną próbkę:*

$$L(\theta; x_1, \ldots, x_n) = L(\theta; \mathbf{x}) = f_n(\mathbf{x}; \theta).$$

Zauważmy, że $L(\theta; \mathbf{x})$ *jest tak naprawdę funkcją* θ*, ponieważ wszystkie parametry* x_i *są ustalone przez próbkę. Metoda MLE polega na znalezieniu takiej wartości* θ*, która maksymalizuje funkcję wiarogodności. Innymi słowy: estymator największej wiarogodności przyjmuje taką wartość, dla której wylosowanie zadanej próbki* $\mathbf{x}$ *staje się najbardziej prawdopodobne. Naszym zadaniem jest zatem znalezienie estymatora*

$$\hat{\theta} : L(\hat{\theta}; \mathbf{x}) = \sup_{\theta} L(\theta; \mathbf{x}).$$

Czasami możliwe jest rozwiązanie tego problemu analitycznie poprzez poszukiwanie maksimum funkcji wiarogodności. W pozostałych przypadkach można uciec do metod numerycznych, jak np. metoda bisekcji, czy metoda stycznych. Czasem jednak metoda ta nie wskazuje jednoznacznie na postać estymatora, lub uzyskanie go może nie być możliwe. Alternatywą może być wtedy metoda momentów.

Metoda momentów *(Method of Moments, MM)* zakłada, że jesteśmy w stanie podać jawne wzory na momenty zmiennej losowej przy użyciu parametrów rozkładu. Jeśli rzeczywiście będziemy w stanie to zrobić, to możemy potraktować zapisane równania jak układ i rozwiązać go ze względu na estymowane parametry.

Algorytm 3.4 (Metoda momentów)**.** *[8] Niech X pochodzi z ustalonego rozkładu opisanego jednoznacznie przez n parametrów. Aby wyestymować wektor parametrów $[\hat{\theta}_1, \ldots, \hat{\theta}_n]$ zapisujemy momenty rozkładu X jako funkcje $g_i(\cdot)$ parametrów rozkładu. Dla większej czytelności niech $m_n = \mathbb{E}[X^n]$.*

$$\begin{aligned} m_1 &= g_1(\theta_1, \ldots, \theta_n) \\ m_2 &= g_2(\theta_1, \ldots, \theta_n) \\ &\vdots \\ m_n &= g_n(\theta_1, \ldots, \theta_n) \end{aligned}$$

Proces estymacji metodą momentów polega na wyestymowaniu $\hat{m}_i = \overline{X^i}$ i rozwiązaniu powyższego układu równań ze względu na estymowane parametry. Tak więc finalnie będziemy mogli obliczyć estymatory jako pewne funkcje $h_i(\cdot)$ wyestymowanych momentów:

$$\begin{aligned} \hat{\theta}_1 &= h_1(\hat{m}_1, \ldots, \hat{m}_n) \\ \hat{\theta}_2 &= h_2(\hat{m}_1, \ldots, \hat{m}_n) \\ &\vdots \\ \hat{\theta}_n &= h_n(\hat{m}_1, \ldots, \hat{m}_n). \end{aligned}$$

Istotne ograniczenie tej metody to rozwiązywalność powstałego układu równań. Nierzadko zdarza się, że albo nie ma on jednoznacznego rozwiązania, albo jest nieliniowy. Wtedy należy użyć momentów innych rzędów, tak aby układ dało się rozwiązać. Trafiamy tu jednak na kolejne ograniczenie - momenty wyższych rzędów mogą nie istnieć, a wtedy nasze estymatory będą bezużyteczne. Z tego samego powodu problematyczna może okazać się estymacja *MM* w klasach rozkładów o wielu parametrach - liczba parametrów może być większa niż liczba dostępnych równań.

Zaprezentowane wyżej metody estymacji w żadnym wypadku nie wyczerpują tematu estymacji. Istnieje dużo więcej sposobów na estymację parametrów rozkładu (np. estymatory Bayesa, czy najmniejszych kwadratów) [8], lecz ich użycie nie jest konieczne w przypadku rozkładów wykorzystywanych w pracy, więc nie będziemy ich tutaj rozważać.

Rozdział 4

Wybrane rozkłady zmiennych losowych

Do modelowania rzeczywistych wielkości losowych wykorzystuje się najczęściej znane klasy rozkładów o dobrze zbadanych właściwościach i potwierdzonej wieloletnią praktyką użyteczności. Wiele z nich ma charakterystyczne dla siebie cechy, które sprawiają że są naturalnymi narzędziami do modelowania określonego zjawiska. Przedstawimy wybrane z nich wraz z ich podstawowymi charakterystykami i sposobami estymacji.

4.1 Rozkład wykładniczy

Definicja 4.1 (Rozkład wykładniczy)**.** [2] Rozkładem wykładniczym o parametrze $\lambda > 0$ nazywamy rozkład $X \sim \mathcal{E}xp(\lambda)$ o dystrybuancie postaci

$$F_X(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

Rozkład wykładniczy z powodu swoich właściwości jest znanym modelem czasu oczekiwania na rzadkie zdarzenia. Parametr λ można tu interpretować jako częstość jego występowania. Zauważmy, że jest on jedynym parametrem rządzącym tym rozkładem. Z tego powodu bardzo łatwo dopasować go do danych, ale kosztem małej elastyczności (nie ma możliwości „poprawy" dobroci dopasowania innym parametrem). Aby ominąć ten problem, czasami stosuje się mieszaninę rozkładów wykładniczych, czyli zmienną która jako dystrybuantę przyjmuje ważoną sumę kilku rozkładów wykładniczych o różnych średnich. Praktyka pokazuje, że taka mieszanina zaledwie dla dwóch składników potrafi dopasować się zaskakująco dobrze [2].

Średnia i wariancja rozkładu wykładniczego to odpowiednio:

$$\mathbb{E}[X] = \frac{1}{\lambda}, \quad \mathbb{V}ar[X] = \frac{1}{\lambda^2}.$$

Widzimy tu ponownie intuicyjną interpretację średniego czasu oczekiwania jako odwrotność częstotliwości występowania modelowanego zdarzenia.

Tak jak wspomnialiśmy wyżej, atrakcyjną cechą tego rozkładu jest łatwość estymacji parametru λ. Zarówno metody *MLE* jak i *MM* zwracają jako estymator odwrotność średniej próbkowej:

$$\hat{\lambda} = \frac{1}{\hat{m}_1}.$$

Równie łatwo symuluje się wartości z rozkładu wykładniczego, chociażby przy pomocy metody odwrotnej dystrybuanty. Jeżeli weźmiemy $U \sim U(0,1)$, to

$$-\frac{1}{\lambda}\ln(U) = X \sim \mathcal{E}xp(\lambda).$$

Bardzo istotną cechą szczególną tego rozkładu jest własność braku pamięci. Matematycznie brak pamięci zapisujemy poprzez prawdopodobieństwo warunkowe:

$$P(X > x + t | X > x) = P(X > t).$$

W praktyce oznacza to, że jeżeli wyczekujemy na zdarzenie które jeszcze się nie wydarzyło, to rozkład czasu oczekiwania na jego wystąpienie nie zależy od momentu rozpoczęcia obserwacji. Własność ta jest ważna w kontekscie tej pracy, ponieważ pozwoli nam w kolejnym rozdziale wyznaczyć przedziały ufności dla funkcji *MEF* opartej o rozkład wykładniczy.

Łatwość estymacji, symulacji oraz prostota konstrukcji rozkładu wykładniczego sprawiają, że jest on podstawowym blokiem budującym wiele modeli procesów stochastycznych (Proces Poissona, czy wiele innych procesów liczących) [11]. Brak pamięci pociąga ponadto za sobą specjalną formę funkcji nadwyżki szkody.

Lemat 4.2 (Funkcja nadwyżki szkody rozkładu wykładniczego)**.** *[2] Jeżeli $X \sim \mathcal{E}xp(\lambda)$, to funkcja nadwyżki szkody przyjmuje postać:*

$$e_X(t) = \frac{1}{\lambda}, \quad t > 0.$$

Dowód. Niech $X \sim Exp(\lambda)$. Wtedy:

$$\mathbb{E}[X - t | X > t] = \frac{\int_t^\infty (x-t)\lambda e^{-\lambda t} dx}{e^{-\lambda t}}.$$

Korzystając z podstawienia $u = x - t$ otrzymujemy:

$$e_X(t) = e^{\lambda t} e^{-\lambda t} \int_0^\infty u\lambda e^{-\lambda u} du = \frac{1}{\lambda},$$

gdzie ostatnią równość zapisujemy znając wartość oczekiwaną rozkładu wykładniczego. □

4.2 Rozkład gamma

Rozkłady gamma pojawiają się w naturalny sposób jako uogólnienie rodziny rozkładów wykładniczych. Zgodność zmiennej losowej X z rozkładem gamma będziemy oznaczać przez $X \sim \mathcal{G}(n, \lambda)$.

Definicja 4.3 (Rozkład gamma)**.** [2] Rozkładem gamma z parametrami $n > 0$, $\lambda > 0$ nazywamy rozkład zmiennej losowej zadany przez gęstość:

$$f_X(x) = \lambda^n x^{n-1} \frac{e^{-\lambda x}}{\Gamma(n)}, \quad x > 0.$$

gdzie $\Gamma(n)$ to funkcja gamma definiowana jako

$$\Gamma(n) \stackrel{def}{=} \int_0^\infty y^{n-1} e^{-y} dy.$$

Zauważmy, że dla $n = 1$ otrzymujemy rozkład wykładniczy, zatem jest on szczególnym przypadkiem tej rodziny rozkładów. Warto wspomnieć również, że dla $n \to \infty$ rozkład gamma dąży do rozkładu normalnego [2]. Parametr n nazywamy parametrem kształtu, a λ parametrem skali. Przy takiej parametryzacji dają się one łatwo interpretować, ponieważ można pokazać (używając chociażby funkcji charakterystycznej), że jeżeli X_i to zmienne losowe niezależne o jednakowym rozkładzie (*iid - independent, identically distributed*) z rozkładu $\mathcal{E}xp(\lambda)$, to:

$$X_1 + X_2 + \cdots + X_n \stackrel{d}{=} Y \sim \mathcal{G}(n, \lambda). \tag{4.1}$$

Zatem możemy interpretować zmienne z rozkładu gamma jako sumę pewnych niezależnych zmiennych losowych o rozkładzie wykładniczym. Z tego powodu rozkład gamma może być wykorzystany do modelowania czasu oczekiwania na n-te zdarzenie. Najprostszym przykładem niech będzie modelowanie niezawodności pewnej maszyny [6], którą przy zepsuciu wymieniamy na nową. Może to być żarówka, opornik, lub inny wybrany element. Jeśli $X \sim \mathcal{E}xp(\lambda)$ modeluje czas jego bezawaryjnego działania, to za pomocą rozkładu $\mathcal{G}(n, \lambda)$ możemy modelować czas oczekiwania na n-tą awarię. Zauważmy, że ponieważ zachodzi własność zadana równością (4.1), to suma dwóch niezależnych zmiennych z rozkładu gamma z tym samem parametrem skali nadal będzie miała rozkład gamma, co więcej - wciąż z tym samym parametrem skali.

Średnia i wariancja rozkładu $\mathcal{G}(n, \lambda)$ wyrażają się jako

$$\mathbb{E}[X] = \frac{n}{\lambda}, \qquad \mathbb{V}ar[X] = \frac{n}{\lambda^2},$$

co jest bezpośrednim następstwem równania (4.1). Estymatory *MLE* dla tego rozkładu nie mają niestety jawnej formy i trzeba uciekać się do metod numerycznych [2]. Na szczęście wykorzystując wzory na pierwsze dwa momenty, możemy za to otrzymać w łatwy sposób estymatory *MM* jako:

$$\hat{n} = \frac{\hat{m}_1^2}{\hat{m}_2 - \hat{m}_1^2},$$
$$\hat{\lambda} = \frac{\hat{m}_1}{\hat{m}_2 - \hat{m}_1^2}.$$

Symulacja zmiennych z rozkładu gamma metodą nie jest możliwa za pomocą metody odwrotnej dystrybuanty, ponieważ ta jest zadana całką. Alternatywnie do symulacji można użyć metody akceptacji-odrzucenia (opierając ją np. na rozkładzie wykładniczym), lecz nie jest to najefektywniejsza metoda. W zastępstwie dostępnych mamy kilka jej usprawnień, oraz zupełnie innych algorytmów na generowanie zgodnie z tym rozkładem [2]. Ponadto jeśli $n \in \mathbb{N}_+$, można skorzystać z równania (4.1), które daje wprost metodę symulacji na podstawie zmiennych o rozkładzie wykładniczym.

Lemat 4.4 (Funkcja nadwyżki szkody rozkładu gamma)**.** *[2] Jeżeli $X \sim \mathcal{G}(n, \lambda)$, to funkcja nadwyżki szkody przyjmuje postać:*

$$e_X(t) = \frac{n}{\lambda} \cdot \frac{1 - CDF_{\mathcal{G}(n+1,\lambda)}(t)}{1 - CDF_{\mathcal{G}(n,\lambda)}(t)} - t = \frac{1}{\lambda}\{1 + o(1)\}, \qquad t > 0.$$

Istotna rzecz którą ukazuje powyższe równanie to fakt, że parametr kształtu nie wpływa na ogon. Dla wystarczająco dużych t realny wpływ będzie miał jedynie parametr skali. Ogon rozkładów gamma będzie zatem przypominał ogon wykładniczy, dla dowolnego parametru kształtu.

4.3 Rozkład Weibulla

Innym sposobem na generalizację rozkładu wykładniczego jest podniesienie zmiennej losowej o tym rozkładzie do pewnej potęgi.

Definicja 4.5 (Rozkład Weibulla)**.** [2] Niech $X \sim \mathcal{E}xp(\lambda)$. Jeżeli rozpatrzymy rozkład zmiennej losowej

$$Y = \sqrt[\tau]{X}, \qquad \tau > 0,$$

to okaże się, że będzie mieć ona dystrybuantę postaci:

$$F_Y(x) = 1 - e^{-\lambda x^\tau}, \qquad x > 0.$$

Rozkład takiej postaci nazywamy rozkładem Weibulla i zapisujemy $Y \sim \mathcal{W}ei(\tau, \lambda)$.

Analogicznie jak w przypadku rozkładu gamma, parametr τ będziemy nazywać parametrem kształtu, a λ parametrem skali. Średnia i wariancja rozkładu Weibulla dane są przez:

$$\mathbb{E}[Y] = \frac{\Gamma\left(1+\frac{1}{\tau}\right)}{\lambda}, \qquad \mathbb{V}ar[Y] = \frac{\Gamma\left(1+\frac{2}{\tau}\right) - \left(\Gamma\left(1+\frac{1}{\tau}\right)\right)^2}{\lambda^2}.$$

Rozkładu Weibulla używa się do modelowania czasów przeżycia, gdy chcemy wprowadzić zależność szansy na zgon od przeżytego wieku. Parametr kształtu τ pozwala nam manipulować intensywnością występowania zgonów w czasie. Jeśli $\tau < 1$, to szansa na zgon będzie wraz z czasem coraz mniejsza. Tak można modelować niezawodność układu w okresie docierania się jego elementów [6] - gdy na początku często psują się wadliwe, a układ robi się coraz odporniejszy na awarię (przez eliminację słabych ogniw). Dla $\tau = 1$ otrzymamy stałą intensywność zgonu (a jednocześnie rozkład sprowadzi się do wykładniczego o parametrze λ). Dla $\tau > 1$ otrzymamy natomiast rosnącą wraz z czasem szansę na zgon (np. model czasu życia człowieka). Parametr kształtu τ wpływa przez to na ciężkość ogonu [2]. Rosnąca szansa zgonu przy $\tau > 1$ powoduje, że rozkład będzie miał lekki ogon. Dla $\tau < 1$ natomiast rozkład Weibulla staje się rozkładem ciężkoogonowym.

Rokład Weibulla ma tę zaletę, że używając metody odwrotnej dystrybuanty można go łatwo wysymulować. Nie da się za to podać jawnych formuł na estymatory *MLE* czy nawet *MM* i trzeba wyliczać je numerycznie.

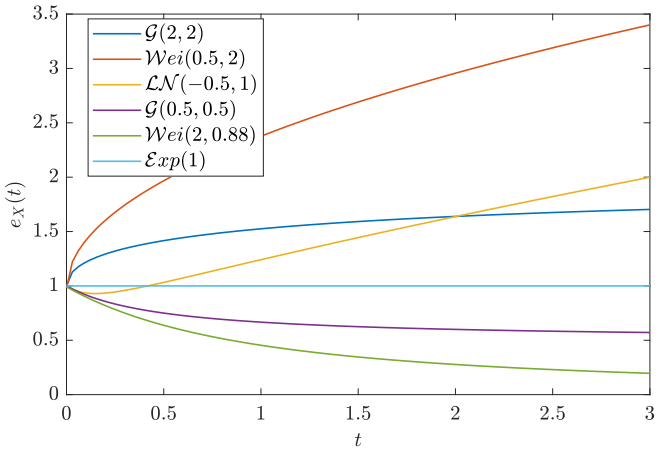
Lemat 4.6 (Funkcja nadwyżki szkody rozkładu Weibulla)**.** *[2] Jeżeli $X \sim \mathcal{W}ei(\tau, \lambda)$, to funkcja nadwyżki szkody przyjmuje postać:*

$$e_X(t) = \frac{\Gamma\left(1+\frac{1}{\tau}\right)}{\lambda^{\frac{1}{\tau}}}\left[1 - \Gamma\left(1+\frac{1}{\tau}, \lambda t^\tau\right)\right]\exp(\lambda t^\tau) - t, \qquad t > 0,$$

gdzie $\Gamma(a,x) = \frac{1}{\Gamma(a)}\int_x^0 t^{a-1}e^{-a}dt$. to niekompletna funkcja gamma.

4.4 Rozkład lognormalny

Ostatnim z omawianych w tej pracy rozkładów jest ciężkoogonowy rokład lognormalny.



Rysunek 4.1: Funkcje nadwyżki szkody wybranych rozkładów o wartości oczekiwanej równej jeden: $\mathcal{G}(2,2), \mathcal{W}ei(0.5,2), \mathcal{LN}(-0.5,1), \mathcal{G}(0.5,0.5), \mathcal{W}ei(2,0.88)$, oraz $\mathcal{E}xp(1)$.

Definicja 4.7 (Rozkład lognormalny)**.** [2] Rozkładem lognormalnym $X \sim \mathcal{LN}(\mu, \sigma^2)$ nazywamy rozkład o gęstości postaci:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2} \frac{(\log x - \mu)^2}{\sigma^2} \right).$$

Rozkład lognormalny jest bardzo popularny w analizie danych rzeczywistych. Posiada on bowiem dodatni nośnik, oraz pozwala dopasować do danych skośność - co nierzadko ukazuje się w realnych danych [7]. Dodatkowo po prostej transformacji rozkład ten zachowuje wiele własności rodziny rozkładów normalnych - co często jest bardzo pożądaną cechą. Rozkład lognormalny pojawia się bowiem, gdy nałożymy eksponentę na zmienną losową z rozkładu normalnego.

Lemat 4.8. *[2] Jeżeli* $N \sim \mathcal{N}(\mu, \sigma^2)$*, to*

$$X = \exp(N) \sim \mathcal{LN}(\mu, \sigma^2).$$

Rozkład lognormalny jest obecny przede wszystkim w matematyce finansowej. Wiele modeli, w tym chociażby najpopularniejszy model wyceny opcji Blacka-Scholesa bazuje bowiem na założeniu, że stopy zwrotu akcji mają właśnie rozkład lognormalny [7]. Mimo, że w praktyce rzadko okazuje się to prawdą (ze względu na dużo cięższe empiryczne ogony), to jednak łatwość estymacji i możliwość otrzymania analitycznych rozwiązań powoduje, że modele te przyjęły się do powszechneg użytku. Dla rozkładu lognormalnego wartość oczekiwana i średnia wyrażają się przez:

$$\mathbb{E}[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \text{ oraz } \quad \mathbb{V}ar[X] = \Big(\exp(\sigma^2) - 1\Big) \exp(2\mu + \sigma^2).$$

Używając powyższych równań można w prosty sposób uzyskać estymatory *MM*, jednak w przypadku tej rozdziny rozkładów dostępne mamy również formuły na estymatory największej wiarogodności:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \log(x_i),$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(\log(x_i) - \hat{\mu})^2.$$

Co ciekawe, rozkład lognormalny jest jednym z rozkładów dla których nie istnieje funkcja generująca momenty [2]. Na szczęście jednak posiada szereg innych zalet, jak na przykład łatwość symulacji. Korzystając z lematu 4.8 widzimy, że wystarczy wygenerować zmienną z rozkładu normalnego (dostępne mamy na to metodę Boxa-Mullera, metodę biegunową i wiele innych [11]) i nałożyć na nią eksponentę.

Lemat 4.9 (Funkcja nadwyżki szkody dla rozkładu lognormalnego)**.** *[2] Dla zmiennej losowej z rozkładu lognormalnego $X \sim \mathcal{LN}(\mu, \sigma)$ funkcja nadwyżki szkody wyraża się wzorem:*

$$e_X(t) = \frac{\exp\left(\mu + \frac{\sigma^2}{2}\right)\left(1 - \Phi\left(\frac{\ln t - \mu - \sigma^2}{\sigma}\right)\right)}{1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)} - t,$$

gdzie $\Phi(x)$ jest dystrybuantą rozkładu standardowego normalnego.

Rozdział 5

Testowanie ogonu rozkładu

Zbierzemy teraz informacje z poprzednich rozdziałów w spójną całość. W kolejnych podrozdziałach przedstawiona zostanie konstrukcja testu zgodności rozkładu próbki z rozkładem wykładniczym o tej samej średniej. W przypadku odrzucenia hipotezy zerowej test będzie klasyfikował próbkę jako ciężkoogonową lub lekkoogonową. Przypomnijmy, że za rozkłady ciężkoogonowe będziemy uznawać zgodnie z definicją 1.10 takie, których ogony zanikają wolniej od rozkładu wykładniczego.

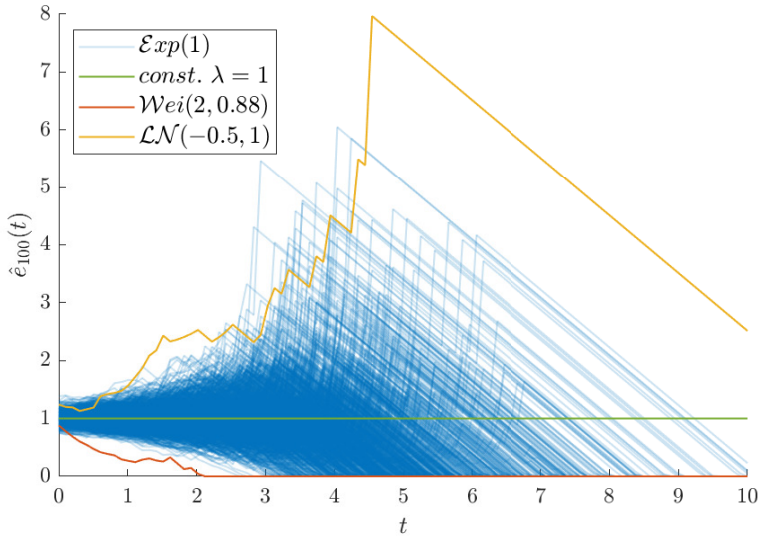
5.1 Konstrukcja testu

W rozdziale drugim mówiliśmy, że funkcja nadwyżki szkody przyjmuje tym większe wartości im ogon próbki jest cięższy i *vice versa*. Z tego powodu naturalnym podejściem do testowania ogonu jest porównywanie wartości jej estymatora, z teoretyczną funkcją *MEF* dopasowanego rozkładu wykładniczego. Jeżeli wartości estymatora są większe, to sygnał, że mamy do czynienia z rozkładem ciężkoogonowym.

Musimy oczywiście pamiętać o tym, że estymator funkcji nadwyżki szkody w każdym punkcie t tak naprawdę jest zmienną losową. Musimy więc mieć jej probabilistyczny opis który pozwoli nam określić w jakich obszarach powinna znaleźć się próbkowa funkcja nadwyżki szkody jeśli stoi za nią rozkład wykładniczy, a które obszary są mniej prawdopodobne. Aby poznać rozkład wartości próbkowego estymatora użyliśmy metody Monte Carlo. Losowaliśmy próbki z rozkładu wykładniczego i rysowaliśmy dla nich trajektorie estymatora. W ten sposób otrzymaliśmy „chmurę" możliwych trajektorii, widoczną na rysunku 5.1. Na tle niebieskiej chmury powstałej z tysiąca realizacji estymatora dla losowych stuelementowych próbek z rozkładu $\mathcal{E}xp(1)$ kolorem zielonym zaznaczono jego teoretyczną funkcję *MEF*.

Na podstawie wysymulowanych trajektorii widzimy dużą stabilność chmury w okolicach zera, po czym możemy obserwować dyfuzję w miarę zwiększania t - zgodnie z wcześniejszymi teoretycznymi rozważaniami. Dodatkowo na rysunek naniesiono innymi kolorami przykładowe realizacje funkcji nadwyżki szkody dla lekkoogonowej wersji rozkładu Weibulla, oraz ciężkoogonowego rozkładu lognormalnego. Już na pierwszy rzut oka odróżniają się one od niebieskich realizacji z rozkładu wykładniczego. Widzimy, że lekkoogonowy rozkład Weibulla osiąga zero zdecydowanie szybciej od nich, a realizacja na podstawie ciężkoogonowej próbki z rozkładu lognormalnego osiąga wielokrotnie większe wartości.

Już w tym momencie można by numerycznie badać przynależność próbki do rodziny rozkładów wykładniczych, symulując dla zadanych t wartości estymatora dla próbek z rodziny rozkładów wykładniczych i sprawdzając jakie w porównaniu z nimi wartości



Rysunek 5.1: Chmura tysiąca realizacji estymatora *MEF* dla stuelementowych próbek z rozkładu $\mathcal{E}xp(1)$, oraz przykładowe realizacje dla rozkładów: $\mathcal{W}ei(2, 0.88)$ oraz $\mathcal{LN}(-0.5, 1)$.

osiągnie estymator badanej próbki.

Pójdziemy jednak o krok dalej i pokażemy, że ten problem posiada rozwiązanie dokładne, teoretyczne. Możemy bowiem znaleźć jawny rozkład estymatora w przypadku gdy próbka pochodzi z rozkładu wykładniczego z parametrem λ.

Twierdzenie 5.1 (Rozkład funkcji nadwyżki szkody dla rozkładu wykładniczego)**.** *Niech* $\mathbf{X} = [X_1, X_2, \ldots, X_n]$ *będą zmiennymi iid z rozkładu* $X_i \sim \mathcal{E}xp(\lambda)$*. Wtedy estymator funkcji nadwyżki szkody tej próbki podlega rozkładowi o dystrybuancie:*

$$P(\hat{e}_n(t) < a) = \sum_{y=0}^{n} \left(CDF_{\mathcal{G}(y,\lambda)}(ay) \binom{n}{y} \frac{\left[1 - CDF_{\mathcal{E}xp(\lambda)}(t)\right]^y}{\left[CDF_{\mathcal{E}xp(\lambda)}(t)\right]^{y-n}} \right), \tag{5.1}$$

gdzie przez $CDF_R(x)$ *oznaczamy dystrybuantę pewnego rozkładu* R *w punkcie* x*.*

Dowód. Weźmy wektor zmiennych losowych $[X_1, X_2, \ldots, X_n]$, gdzie X_i są *iid* o rozkładzie $\mathcal{E}xp(\lambda)$. Rozważymy funkcję nadwyżki szkody

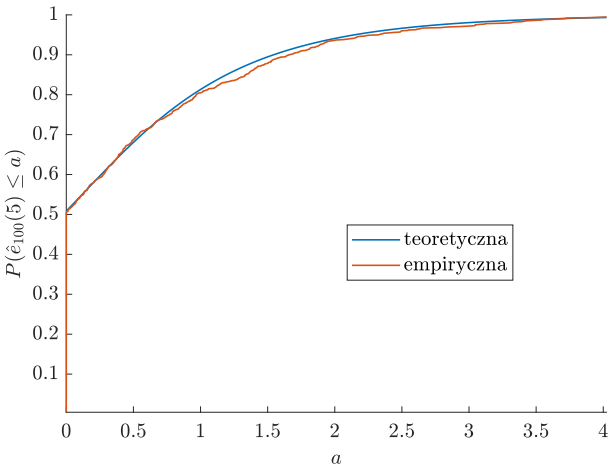
$$\hat{e}_n(t) = \frac{\sum_{X_i > t} X_i}{\#\{i : X_i > t\}} - t,$$

pod kątem partycji zadanej liczbą obserwacji spełniających warunek $X_i > t$. Ze wzoru na prawdopodobieństwo całkowite możemy otrzymać:

$$P(\hat{e}_n(t) < a) = \sum_{y=0}^{n} P\left(\frac{\sum_{X_i > t} X_i - t}{\#\{i : X_i > t\}} < a \middle| \#\{i \colon X_i > t\} = y \right) P(\#\{i \colon X_i > t\} = y). \tag{5.2}$$

Dla czytelności wprowadźmy oznaczenia:

$$M := \#\{i \colon X_i > t\}, \text{ oraz } \quad N := \sum_{\{X_i > t\}} (X_i - t).$$



Rysunek 5.2: Porównanie dystrybuant: empirycznej i teoretycznej estymatora *MEF* w punkcie $t = 5$ dla stuelementowej próby losowej z rozkładu $\mathcal{E}xp(1)$.

Zauważmy teraz, że zmienna losowa M może być interpretowana jako „liczba sukcesów" - jeśli za sukces przyjmiemy, że realizacja z rozkładu $\mathcal{E}xp(\lambda)$ jest większa od t. Ma zatem ona rozkład dwumianowy z liczbą prób n, czyli:

$$P(M = m) = \binom{n}{m}\Big(1 - CDF_{\mathcal{E}xp(\lambda)}(t)\Big)^m\Big(CDF_{\mathcal{E}xp(\lambda)}(t)\Big)^{n-m}.$$

Zauważmy ponadto, że dla $X_i \sim \mathcal{E}xp(\lambda)$ dzięki własności braku pamięci rozkładu wykładniczego mamy $(X_i - t)|(X_i > t) \stackrel{d}{=} X_i$, więc co do rozkładu zachodzi równość:

$$\sum_{\{X_i > t\}} (X_i - t)|(X_i > t) \stackrel{d}{=} \sum_{i=0}^{n} X_i.$$

Z własności (4.1) rozkładu gamma otrzymujemy zatem

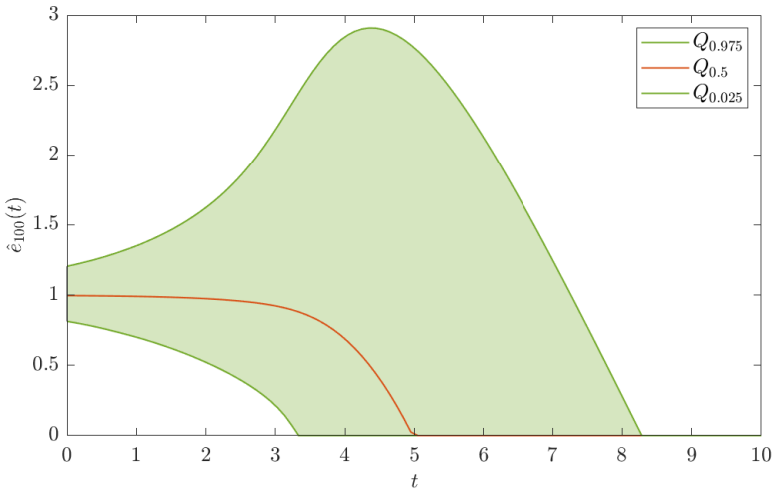
$$N \stackrel{d}{=} \sum_{i=0}^{n} X_i \sim \mathcal{G}(n, \lambda).$$

Robiąc użytek z powyższych faktów, możemy przepisać równanie 5.2 do formy

$$\begin{aligned} P(\hat{e}_n(t) < a) &= \sum_{y=0}^{n} P\left(\frac{\sum_{X_i > t} X_i - t}{\#\{i : X_i > t\}} < a \Big| \#\{i \colon X_i > t\} = y\right) P(\#\{i \colon X_i > t\} = y) = \\ &= \sum_{y=0}^{n} P\left(\frac{N}{y} < a \Big| M = y\right) P(M = y) = \\ &= \sum_{y=0}^{n} \left(\mathrm{CDF}_{\mathcal{G}(y,\lambda)}(ay)\binom{n}{y}\frac{\Big[1 - \mathrm{CDF}_{\mathcal{E}xp(\lambda)}(t)\Big]^y}{\Big[\mathrm{CDF}_{\mathcal{E}xp(\lambda)}(t)\Big]^{y-n}}\right). \end{aligned}$$

□

Otrzymujemy w ten sposób rozkład wartości funkcji *MEF* opartej o n-elementowy losowy wektor z rozkładu $\mathcal{E}xp(\lambda)$, przy ustalonym argumencie t. Symulacyjnie została sprawdzona zbieżność wyniku teoretycznego do empirycznego rozkładu wartości estymatora.



Rysunek 5.3: Numerycznie wyznaczone kwantyle $Q_{0.025}$ oraz $Q_{0.975}$ tworzące przedziały ufności dla $\hat{e}_{100}(t)$ na poziomie ufności 95% - przypadek rozkładu $\mathcal{E}xp(1)$.

Zbieżność dystrybuant dla przykładowych wartości parametrów $\lambda = 1$, $n = 100$ oraz $t = 5$ została zaprezentowana na rysunku 5.2.

Zwróćmy uwagę na fakt, że nie jest to dystrybuanta ciągła, ponieważ istnieje niezerowe prawdopodobieństwo, że zmienna przyjmie wartość zero. Empirycznie jest to związane z możliwością, że w niektórych stuelementowych próbkach wszystkie wartości będą mniejsze od $t = 5$. To właśnie dzięki przedefiniowaniu funkcji *MEF* (definicja 2.3) otrzymana funkcja może osiągnąć zero. Gdyby nie zmiana definicji, moglibyśmy co najwyżej przejść na rozkłady warunkowe[1].

Skok w otrzymanej dystrybuancie równy jest zatem prawdopodobieństwu, że realizacja funkcji nadwyżki szkody może spaść do zera dla argumentów mniejszych od t:

$$P\Big(\inf_{\tau>0}\{\hat{e}_n(\tau) = 0\} < t\Big) = P(\max \mathbf{X} < t).$$

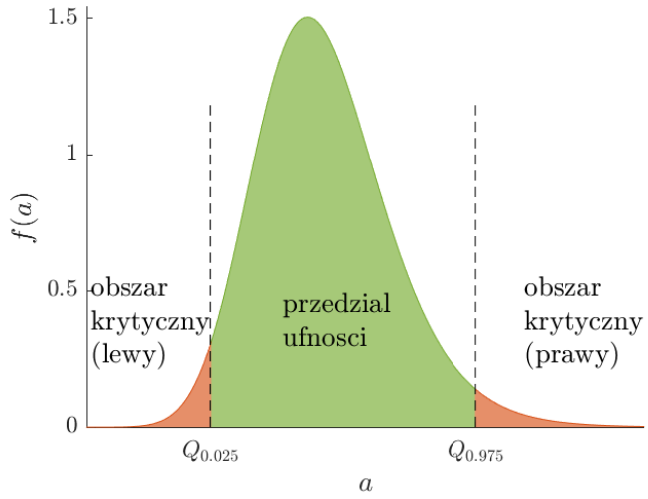
Przy dowolnych wartościach n, α w miarę zmniejszania parametru t zmniejsza się również wyżej wspomniany skok dystrybuanty, ponieważ $P(\max \mathbf{X} < t)$ maleje do zera. W rezultacie, dla wystarczająco małych wartości t dystrybuanta będzie mieć na tyle mały skok, że będzie dawać się przybliżyć poprzez dystrybuantę ciągłą.

Wykorzystując twierdzenie 5.1 jesteśmy w stanie określić na dowolnym poziomie ufności przedziały ufności dla funkcji *MEF*. Przykładowo możemy otrzymać przedziały ufności na poziomie ufności 95% dla rozkładu $\hat{e}_n(t)$ rozwiązując równania:

$$P(\hat{e}_n(t) < Q_{0.025}) = 0.025, \text{ oraz } \quad P(\hat{e}_n(t) < Q_{0.975}) = 0.975.$$

Numeryczne rozwiązanie tych równań dla $n = 100$ można obejrzeć na rysunku 5.3. Możemy na ich podstawie skonstruować test, który na zadanym poziomie istotności odrzuci próbkę jeżeli statystyka testowa (tu: empiryczna funkcja nadwyżki szkody) trafi poza obliczony przedział.

[1] Wymagałoby to warunkowania wartości funkcji nadwyżki szkody poprzez jej istnienie



Rysunek 5.4: Gęstość $f(a)$ rozkładu $\hat{e}_{100}(2)$ oraz jego 95% przedziały ufności (kolor zielony) dla próbki z rozkładu $\mathcal{E}xp(1)$.

5.2 Testowanie zgodności z rozkładem wykładniczym

Test 5.2 (Test zgodności z rozkładem wykładniczym)**.** *Niech* $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ *będzie próbką, którą chcemy przetestować pod kątem zgodności z rozkładem wykładniczego. W proponowanym teście, za hipotezę zerową będziemy przyjmować zatem* H_0 *: „Próbka ma rozkład wykładniczy", a testować będziemy przeciwko* H_1 *: „Próbka ma inny rozkład niż wykładniczy". Oznaczmy przez* α *wybrany poziom istotności testu. Najpierw należy wyestymować parametr* λ *na podstawie próbki. Estymujemy go metodą największej wiarogodności:*

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^{-1}.$$

Za pomocą twierdzenia 5.1 tworzymy dwustronny przedział ufności $(Q_{\frac{\alpha}{2}}, Q_{1-\frac{\alpha}{2}})$ *na wybranym poziomie ufności, przy wyestymowanym* $\hat{\lambda}$ *i przy wybranym parametrze* t *(argument empirycznej funkcji nadwyżki szkody). W tym celu szukamy kwantyli* Q_p *rozkładu estymatora funkcji nadwyżki szkody rozwiązując równania:*

$$P\left(\hat{e}_n(t) < Q_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}, \quad oraz \quad P\left(\hat{e}_n(t) < Q_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Za statystykę testową Z *przyjmować będziemy empiryczną funkcję nadwyżki szkody 2.2 obliczaną w oparciu o testową próbkę, przy wybranym argumencie* t*:*

$$Z = \hat{e}_n(t) = \frac{\sum_{\{x_i > t\}} x_i}{\#\{i : x_i > t\}} - t.$$

Jeżeli statystyka testowa znajdzie się wewnątrz przedziału ufności, czyli $Z \in (Q_{\frac{\alpha}{2}}, Q_{1-\frac{\alpha}{2}})$*, to nie mamy podstaw aby odrzucać hipotezę zerową. W przeciwnym wypadku powinniśmy odrzucić hipotezę* H_0 *na rzecz* H_1*.*

Powyższy test wykorzystuje analityczny rozkład funkcji nadwyżki szkody aby sprawdzić czy rozkład pochodzi z rodziny wykładniczych. Wiemy jednak również, że rozkład wykładniczy stanowi przejście pomiędzy rozkładami ciężkoogonowymi a lekkoogonowymi.

Jeżeli zatem algorytm 5.2 odrzuci hipotezę przynależności do rodziny rozkładów wykładniczych, to pozostają nam tylko dwie alternatywy: rozkład musi mieć albo lekki, albo ciężki ogon. Używając przy testowaniu dwustronnego przedziału ufności tworzymy dwa rozłączne obszary krytyczne (zaprezentowane na rysunku 5.4 kolorem czerwonym). Jeżeli statystyka testowa znajdzie się w którymś z nich, to będziemy odrzucać hipotezę zerową. Dodatkowo zgodnie ze wcześniejszą uwagą - dla rozkładów o ogonie cięższym od wykładniczego funkcja *MEF* przyjmuje większe wartości. Z tego powodu trafienie statystyki testowej w prawy obszar krytyczny sugerować będzie ciężkoogonowość rozkładu z którego pochodzi badana próbka. Również w drugą stronę - lekkoogonowość próbki zwiększa szanse trafienia statystyki testowej w lewy obszar krytyczny. W ten sposób możemy wzbogacić test 5.2 o regułę klasyfikującą ogon rozkładu w przypadku odrzucenia hipotezy zerowej.

Możemy również zmodyfikować algorytm, aby wprost testować czy rozkład ma ciężki (alternatywnie lekki) ogon. Możemy użyć jednostronnych przedziałów ufności w celu testowania hipotezy zerowej H_0 : próbka ma ogon cięższy niż wykładniczy przeciw hipotezie alternatywnej H_1 : próbka ma ogon wykładniczy lub lżejszy od ogonu wykładniczego[2].

5.3 Parametry testu a jego skuteczność

Formułując procedurę wykonywania testu 5.2 pominęliśmy kwestię doboru parametrów: n, α, λ, t, którym należy nadać odpowiednie wartości przed rozpoczęciem testowania.

Parametr n jest długością badanej próbki, a więc jest jednoznacznie określony. Równie jasny w doborze powinien być również poziom istotności α za pomocą którego zmieniamy prawdopodobieństwo popełnienia błędu pierwszego rodzaju. W praktyce przyjęło się ustalenie $\alpha = 0.05,$ lub $\alpha = 0.01$, co wprost odpowiada prawdopodobieństwu odrzucenia hipotezy zerowej mimo jej prawdziwości.

Jak wspomnieliśmy wcześniej, mamy również dowolność wyboru rodzaju przedziału ufności. Dwustronny przedział ufności jest bardziej naturalny przy testowaniu zgodności zaniku ogonu próbki z zanikiem wykładniczym, ponieważ odcina najmniej prawdopodobne obserwacje równomiernie z prawego i lewego ogona próbki. Jednostronne przedziały ufności można natomiast wykorzystać do testowania konkretnej hipotezy na temat ciężkości ogonu.

Parametr λ rządzi rozkładem wykładniczym, który służy do otrzymania hipotetycznego rozkładu estymatora *MEF*. Tutaj również nie powinno budzić wątpliwości wykorzystanie metody największej wiarogodności w celu estymacji tego parametru:

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^{-1}.$$

Natomiast ostatnim i najmniej jasnym w doborze parametrem jest t, dlatego poświęcimy najwięcej czasu na jego analizę. Zależy od niego kształt dystrybuanty estymatora funkcji nadwyżki szkody, a co za tym idzie szerokość i położenie przedziału ufności, a w konsekwencji skuteczność testu.

Rozpocznijmy od dolnego ograniczenia: ponieważ badamy dodatnie ogony, to wolno nam skupić się na $t \geqslant 0$. Wiemy również, że t jest granicą; punktem odcięcia który oddziela obserwacje istotne w obliczaniu estymatora od tych które nie są brane pod uwagę. Granicznie jeśli przyjmiemy $t = 0$, to nie odetniemy żadnej obserwacji, a funkcja *MEF* będzie próbkową średnią. Wtedy dwa rozkłady unormowane do tej samej średniej byłyby nie do odróżnienia przez test, choćby miały skrajnie różne ciężkości ogonów. Jeśli natomiast

[2]Testowanie lekkiego ogonu przeprowadzimy odpowiednio zmieniając ciężkość w obu hipotezach

przypiszemy t zbyt dużą wartość ($t \geqslant \max\{\mathbf{x}\}$), to estymator przyjmie wartość zero i nie powie nic istotnego na temat próbki, nie licząc oczywiście tego, że $\max\{\mathbf{x}\} \leqslant t$. W ten sposób testowalibyśmy jednak maksimum próbkowe. Weszlibyśmy w teorię wartości ekstremalnych (*Generalized Extreme Value Theory, GEV*) która jest pod tym względem dobrze zbadana i opisana [7][10]. Spróbujemy zatem zbadać co dzieje się na przedziale $t \in (0, \max\{\mathbf{x}\})$.

Jak wspominaliśmy wcześniej rozkład statystyki testowej nie jest ciągły; tylko dla wystarczająco małych t daje się przez taki przybliżyć. To wprowadza istotne ograniczenie górne na wartości t jakich możemy używać. Problematyczne będzie bowiem przeprowadzenie testu na dowolnym poziomie istotności α, gdy statystyka testowa będzie miała rozkład dyskretny, lub (jak w naszym przypadku) mieszany. Ograniczenie wprowadzone przez to wygląda następująco:

$$P(\hat{e}_n(t) = 0) < \frac{\alpha}{2}. \tag{5.3}$$

Dla spełniających je t da się jednoznacznie wskazać dolny kwantyla rzędu $\frac{\alpha}{2}$. Gdy ten warunek nie jest spełniony, test będzie miał problem nie tylko z prawidłowym ustaleniem przedziałów ufności, ale i z wykrywaniem lekkoogonowych cech rozkładu. Tyczy się to również przypadku użycia w teście jednostronnego przedziału ufności postaci $(Q_{0.025}, \infty)$, choć warunek (5.3) będzie miał wtedy nieco inną postać.

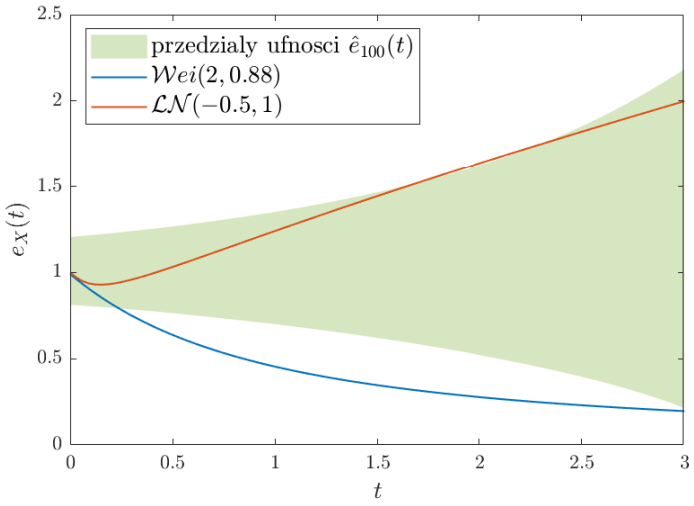
Załóżmy dalej, że wybieramy t spełniające wyżej wymienione warunki. Moglibyśmy wtedy powiedzieć, że wybór t nie ma znaczenia. Test przecież teoretycznie działa dla dowolnie wybranej jego wartości, odrzucając hipotezę zerową jeśli statystyka testowa znajdzie się poza przedziałami ufności.

Zauważmy jednak, że podczas wykonywania testu pracujemy na rozkładach dwóch zmiennych losowych. Pierwszy z nich to hipotetyczny, zakładany z góry rozkład jaki powinna mieć statystyka testowa, określany przez założenia w momencie formułowania hipotezy zerowej. Drugi z nich jest natomiast związany z *rzeczywistym* rozkładem z którego pochodzi próbka, co pociąga za sobą pewien *rzeczywisty* rozkład jaki posiada statystyka testowa.

Dla dobrego testu statystycznego rozkłady te muszą pokrywać się przy prawdziwości hipotezy zerowej, a w wypadku gdy ta nie zachodzi rozkłady powinny od siebie odbiegać. Test jest tym lepszy im mniej się one pokrywają, ponieważ prawdopodobieństwa popełnienia błędu pierwszego i drugiego rodzaju maleją. Nie można niestety oszacować tego stopnia pokrycia z góry, ponieważ każdy sposób fałszu w hipotezie zerowej może prowadzić do innego realnego rozkładu statystyki testowej. Możemy jednak dobrać wartość parametru t w taki sposób, żeby przy pewnym ustalonym scenariuszu jak najbardziej oddalić od siebie rozkłady: hipotetyczny i realny statystyki testowej.

Zobaczmy graficznie tę sytuację na rysunku 5.5. Przedstawia on przedziały ufności na poziomie ufności 95% dla statystyki testowej przy prawdziwej hipotezie zerowej (kolor zielony). Innymi kolorami naniesiono natomiast teoretyczne funkcje nadwyżki szkody innych, przykładowych rozkładów. Ustalając konkretny parametr t, możemy obserwować jak będą wyglądały przedziały ufności estymatora dla rozkładu wykładniczego, oraz jak odległa od niego jest wartość teoretyczna funkcji nadwyżki szkody obcego rozkładu.

Widzimy na przykład, że dla rozkładu $\mathcal{W}ei(2, 0.88)$ wybranie $t \in (1, 1.5)$ będzie rozsądniejsze niż $t \in (2.5, 3)$, ponieważ daje to większą szansę, że fałszywa hipoteza zerowa zostanie prawidłowo odrzucona. Przykładowo jednak dla rozkładu $\mathcal{LN}(-0.5, 1)$ mamy już zupełnie odwrotną sytuację. Tutaj wybranie $t \in (1, 1.5)$ sprawia, że nawet przy niespełnieniu hipotezy zerowej rozkłady w sporym stopniu pokryją się i odrzucenie fałszywej hipotezy zerowej może być problematyczne. Jeśli wybierzemy natomiast $t \in (2, 2.5)$, to



Rysunek 5.5: Funkcje *MEF* rozkładów: $\mathcal{W}ei(2, 0.88)$, oraz $\mathcal{LN}(-0.5, 1)$ na tle przedziałów ufności na poziomie ufności 95% dla estymatora $\hat{e}_{100}(t)$ rozkładu $\mathcal{E}xp(1)$.

odrzucenie takiej hipotezy będzie relatywnie łatwiejsze z powodu mniejszego pokrycia funkcji gęstości rozkładu hipotetycznego i realnego.

Skoro naszym celem jest klasyfikacja ogonu rozkładu uzasadniony wydaje się taki wybór t, który pozwala na największą moc testu - czyli jak największe prawdopodobieństwo na to, że fałszywa hipoteza zerowa zostanie odrzucona. Jeśli nie odrzucimy hipotezy zerowej, to nie będziemy mieli bowiem informacji o wykrytym zachowaniu ogonu.

Oznaczmy przez $F_{R_t}(x)$ dystrybuantę realnego rozkładu (zależnej od t) statystyki testowej $M(t) = \hat{e}_n(t)$, a przez $F_H(x)$ dystrybuantę rozkładu statystyki testowej przy prawdziwości hipotezy zerowej. Niech ponadto Q_p oznacza kwantyl rzędu p rozkładu $F_H(x)$. Problem wyboru parametru t można wtedy przedstawić jako:

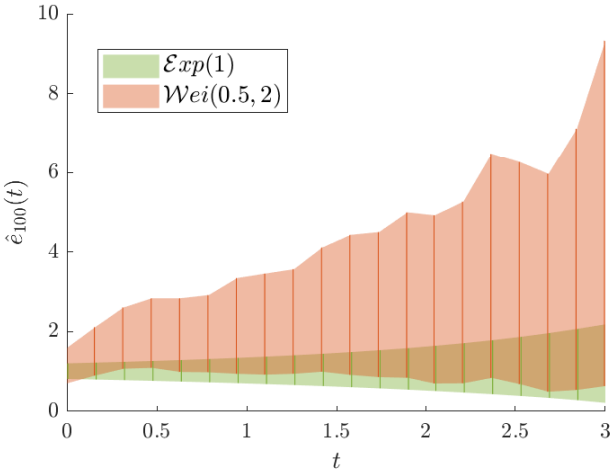
$$t\colon P\Big(M(t) \notin (Q_{\frac{\alpha}{2}}, Q_{1-\frac{\alpha}{2}})\Big) = \max_{\tau>0}\Big(P(M(\tau) < Q_{\frac{\alpha}{2}}) + P(Q_{1-\frac{\alpha}{2}} < M(\tau))\Big) = \\ = \max_{\tau>0}\Big(1 + F_{R_\tau}(Q_{\frac{\alpha}{2}}) - F_{R_\tau}(Q_{1-\frac{\alpha}{2}})\Big).$$

Podczas testowania próbki nie znamy jednak $F_{R_t}(x)$, więc nie możemy rozwiązać tego problemu analitycznie ani numerycznie. Metodologia proponowana w tej pracy zamiast badania całego rozkładu opiera się badaniu odległości dwóch przedziałów ufności: teoretycznego dla estymatora *MEF* rozkładu $\mathcal{E}xp(\hat{\lambda})$, oraz wysymulowanego z realnych danych.

Nie chcemy zakładać, że z góry znamy rozkład z którego pochodzi próbka - to nie miałoby sensu, bo możnaby od razu wyestymować parametry i powiedzieć na ich podstawie czy mamy do czynienia z ciężkoogonowym rozkładem. Skupimy się na sytuacji trudniejszej ale i częściej spotykanej w praktyce. Będziemy mieć pojedynczą próbkę obserwacji i brak informacji o rozkładzie z którego została wylosowana.

Aby wybrać parametr t poprzez maksymalizację mocy testu potrzebujemy jednak mieć więcej niż jedną próbkę. Należy przecież stworzyć przedziały ufności dla próbkowej funkcji nadwyżki szkody. Skorzystamy zatem z procedury *bootstrap* [8], która polega na tworzeniu wariacji z powtórzeniami dostępnej próbki.

Majac próbkę $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ losujemy z niej (z powtórzeniami) n wartości, które stworzą nową próbkę $\mathbf{x}_{(1)}$. Wykonujemy ten krok wiele razy, oznaczając nowo tworzone wariacje jako $\mathbf{x}_{(2)}, \mathbf{x}_{(3)}, \ldots$ Możemy następnie dla każdej wariacji obliczyć estymator funkcji



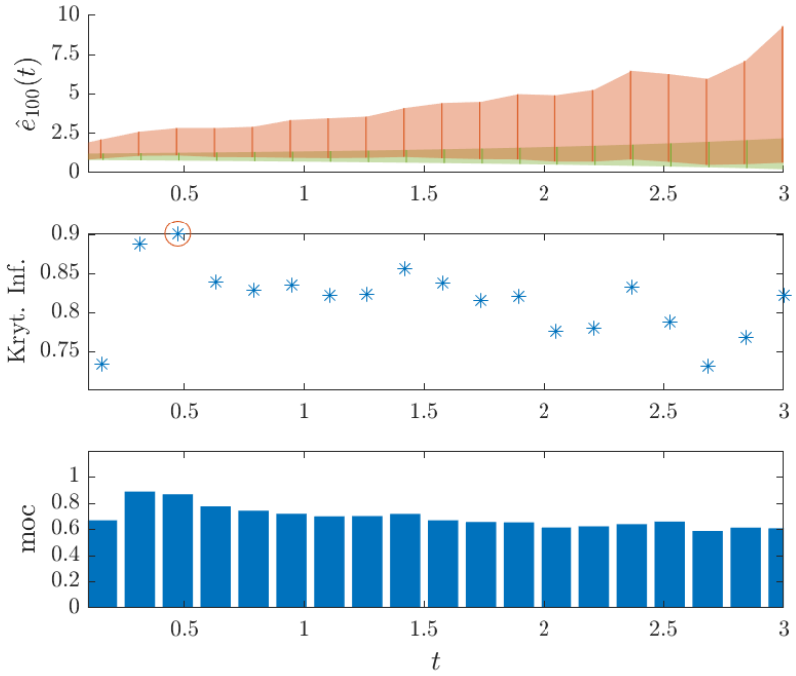
Rysunek 5.6: Przedziały ufności estymatora $\hat{e}_{100}(t)$ na poziomie ufności 95%: teoretyczne w przypadku rozkładu $\mathcal{E}xp(1)$, oraz otrzymane metodą *bootstrap* dla rozkładu $\mathcal{W}ei(0.5, 2)$.

nadwyżki szkody, a ze zbioru tych estymatorów otrzymać empiryczne przedziały ufności dla wybranego t na wybranym poziomie ufności α.

Dla dowolnego t otrzymamy w ten sposób parę przedziałów ufności. Wynik takiej procedury widoczny jest na rysunku 5.6. Kolorem zielonym zaznaczono przedziały ufności na poziomie ufności 95% dla estymatora funkcji nadwyżki szkody stuelementowej próbki z rozkładu $\mathcal{E}xp(1)$, a kolorem czerwonym takie same przedziały ufności w przypadku rozkładu $\mathcal{W}ei(0.5, 1)$. Dla tego ostatniego przedziały ufności zostały ustalone na podstawie jednej oryginalnej próbki i 1000 powtórzeń *bootstrap*. Cały proces jest kosztowny obliczeniowo, dlatego przyjęta została niezbyt gęsta siatka równomiernie rozmieszczonych punktów w których przeprowadzamy tę procedurę. Na rysunku są one widoczne jako pionowe pręgi. Brzegi powstałych przedziałów przedziałów ufności zostały zinterpolowane w celu stworzenia widocznego obszaru. Dla wybranych t mamy zatem dwa przedziały ufności: jeden przedział teoretyczny, a jeden empiryczny na podstawie próbki. Zwróćmy uwagę na fakt, że w ilustrowanym przykładzie wybór t jest zbyteczny, ponieważ już po rzucie oka na rysunek 5.6 doskonale widać jaki ogon ma badana próbka.

Mając dostępne różne wartości t do wyboru, należy stworzyć pewne kryterium informacyjne pozwalające sprawdzić jak „dobra" jest każda para przedziałów. Postulowane kryterium powinno faworyzować pary przedziałów, które przy zadanej próbce maksymalizują moc testu - ponieważ jak pisaliśmy wyżej to zwiększa szansę na klasyfikację ogonu próbki. Jedna rzecz wydaje się być w takim razie jasna: będziemy preferować t dla których przedziały ufności są rozłączne. Nakładanie się przedziałów ufności oznacza bowiem istotne prawdopodobieństwo popełnienia błędu drugiego rodzaju. Równie jasne jest, że jeżeli mamy więcej niż jedno t dla którego przedziały ufności są rozłączne, to powinniśmy wybrać takie, dla którego przedziały są od siebie najbardziej oddalone, aby tym bardziej zniwelować prawdopodobieństwo niesklasyfikowania ogonu próbki.

W ostatnim przypadku - przy częściowym pokryciu przedziałów ufności proponujemy zbadanie procentowego stopnia tego pokrycia. Chcemy żeby nasza statystyka testowa (o czerwonym *realnym* rozkładzie - nie istotne czy zgodnym z hipotezą zerową) miała jak największą szansę na trafienie poza przedział ufności *hipotetycznego* rozkładu. W tym świetle para przedziałów będzie tym lepsza, im większy procent empirycznego (*bootstrap*) przedziału ufności będzie znajdował się poza teoretycznym (*hipotetycznym*) przedziałem



Rysunek 5.7: Górny panel: rysunek 5.6. Środkowy panel: wartości kryterium informacyjnego dla zadanych t. Dolny panel: symulacyjnie otrzymana moc testu przy zadanym t.

ufności.

Wynik działania opisanej wyżej procedury można zobaczyć na rysunku 5.7. Poddaliśmy przedziały ufności z rysunku 5.6 komputerowemu programowi wybierającemu t zgodnie z opisaną wyżej procedurą. Na górnym panelu rysunku widzimy że rozkłady odbiegają od siebie - otrzymujemy sporo przedziałów niemal rozłącznych, co odzwierciedlone zostało na środkowej części rysunku przez wartości kryterium bliskie jedynce. Algorytm wybrał zatem spośród badanych t takie, dla którego przedziały mają największy rozrzut i oznaczył odpowiadającą mu wartość kryterium informacyjnego czerwoną otoczką. Dodatkowo aby sprawdzić symulacyjnie moc testu dla każdego t symulowaliśmy 1000 próbek na podstawie których algorytm 5.2 odrzucał lub akceptował hipotezę zerową. Moc testu (procent odrzuconych próbek) przedstawiono na diagramie słupkowym na dolnym panelu rysunku 5.7. Wybrana wartość $t \approx 0.95$ ma maksymalizować procent próbek w których dochodzi do odrzucenia hipotezy zerowej. Istotnie, na dolnym panelu rysunku widzimy, że dla takiej wartości parametru uzyskujemy przy zadanej próbce moc testu na poziomie około 96%. Istotna rzecz której nie widać na rysunku to fakt, że każde odrzucenie hipotezy zerowej [3] było spowodowane trafieniem statystyki testowej w obszar krytyczny odpowiadający ciężkiemu ogonowi.

[3] Wyłączając najmniejsze badane t.

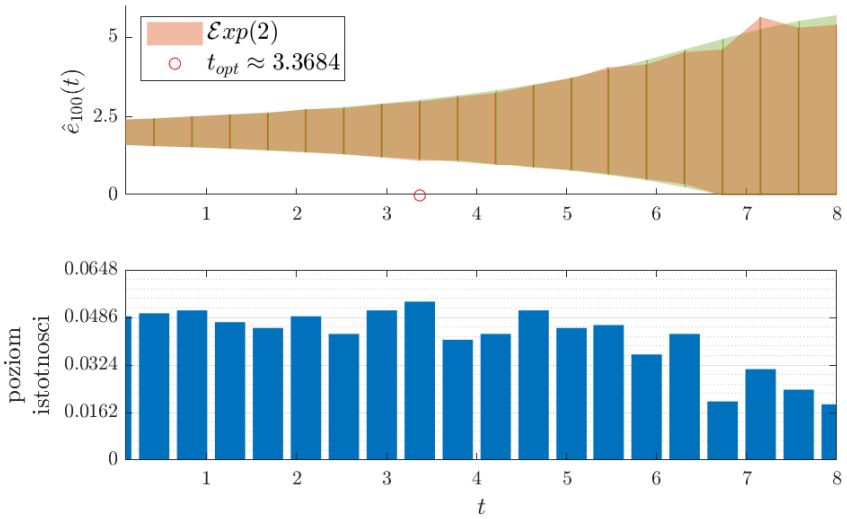
Rozdział 6

Zastosowanie testu do wybranych rozkładów oraz danych rzeczywistych

Zaprezentowany powyżej algorytm pokazuje statystyczną metodę klasyfikacji ogonu próbki. W tym rozdziale zbadamy jak test zachowuje się w rodzinach rozkładów: wykładniczych, Weibulla, gamma oraz lognormalnych. We wszystkich przykładach będziemy testować z użyciem dwustronnych przedziałów ufności na poziomie ufności 95%.

6.1 Testowanie w wybranych klasach rozkładów

6.1.1 Rozkłady wykładnicze



Rysunek 6.1: Algorytm wyboru optymalnego parametru t dla próbki z rozkładu $\mathcal{E}xp(2)$. Górny panel: hipotetyczne (zielone) i empiryczne (czerwone) przedziały ufności estymatora funkcji nadwyżki szkody na poziomie ufności 95%. Dolny panel: symulacyjnie wyznaczony poziom istotności testu.

Na początku sprawdzimy zachowanie testu w przypadku prawdziwej hipotezy zerowej. Do przeprowadzenia testu zilustrowanego na rysunku 6.1 użyliśmy stuelementowej próbki z rozkładu $\mathcal{E}xp(2)$. W celu ustalenia empirycznych przedziałów ufności estymatora nie

używaliśmy tu metody *bootstrap*, lecz generowaliśmy próbki bezpośrednio z generatora zmiennych losowych $\mathcal{E}xp(\hat{\lambda})$. Wynik działania testu możemy obserwować na rysunku 6.1. Górny panel rysunku demonstruje zbieżność hipotetycznych i empirycznych przedziałów ufności na całym rozpatrywanym spektrum parametru t. Dolny panel natomiast ilustruje poziom istotności testu obliczony symulacyjnie jako procent próbek dla których została odrzucona hipoteza zerowa spośród 1000 próbek losowych o rozkładzie $\mathcal{E}xp(2)$. Przy prawdziwej hipotezie zerowej (tak właśnie jak w przypadku $X \sim \mathcal{E}xp(2)$) odrzucane powinno być $\alpha = 5\%$ realizacji. Istotnie na dolnym panelu widzimy, że test działa prawidłowo, ponieważ otrzymujemy poziomy istotności w okolicach 0.05. Zwróćmy uwagę na spadek poziomu istotności dla czterech brzegowych wartości t po prawej stronie. Powodem jest niemożność zdefiniowania tam kwantyla rzędu $Q_{0.05}$ dla teoretycznego rozkładu estymatora - o czym wspominaliśmy w części pracy poświęconej wyborze parametru t. W związku z tym otrzymujemy dla takich t tylko jeden przedział krytyczny, a poziom istotności spada o połowę.

W celu zbadania poziomu istotności testu dla rodziny rozkładów wykładniczych przeprowadziliśmy test na stuelementowych próbkach z rozkładu $\mathcal{E}xp(\lambda)$ dla 20 różnych wartości parametru λ. Dla każdej wartości λ przeprowadzono 150 powtórzeń testu zapisując za każdym razem osiągniety symulacyjnie poziom istotności odpowiadający wybranej optymalnej wartości t. Tabela 6.1 przedstawia wartości otrzymanego symulacyjnie poziomu istotności $\tilde{\alpha}$ testu dla badanych wartości λ. Zgodnie z oczekiwaniami prawdziwość hipotezy zerowej powoduje, że dla wszystkich wartości λ otrzymujemy symulacyjny poziom istotności testu oscylujący w okolicach 5%. Powyższe wyniki w rodzinie rozkładów wykładniczych (przy prawdziwej hipotezie zerowej) upewniają nas co do poprawności działania testu.

$\tilde{\alpha}$	0.048	0.047	0.051	0.046	0.054	0.057	0.043	0.046	0.051	0.043
λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\tilde{\alpha}$	0.045	0.053	0.050	0.048	0.046	0.058	0.041	0.054	0.034	0.051
λ	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2

Tabela 6.1: Wartości uśrednionego symulacyjnego poziomu istotności testu $\tilde{\alpha}$ dla stuelementowej próby losowej z rozkładu $\mathcal{E}xp(\lambda)$. *Źródło: opracowanie własne.*

6.1.2 Rozkłady gamma

Rodzina rozkładów gamma charakteryzowana jest przez dwa parametry. I choć jak mówiliśmy wcześniej parametr kształtu nie wpływa na ciężkość ogonu [2], to wpływa na kształt funkcji nadwyżki szkody. Będziemy zatem badać zależność mocy testu $\tilde{\theta}$ od obu parametrów tego rozkładu.

Wybraliśmy po 5 przykładowych wartości dla parametrów n oraz λ z których stworzyliśmy możliwe pary (n, λ). Dla każdego punktu tak stworzonej siatki przeprowadziliśmy 150 powtórzeń testu dla próbek o długości stu elementów z odpowiednich rozkładów $\mathcal{G}(n, \lambda)$. Podobnie jak wcześniej otrzymane symulacyjnie moce testu zostały uśrednione dla każdego punktu. Wyniki eksperymentu są widoczne w tabeli 6.2.

Widzimy, że test ma problem z klasyfikacją ogonów rozkładów gamma, ponieważ symulacyjnie jego moc bardzo mocno się waha. Najmniejszą moc test wykazuje gdy $n = 1$ - czyli w przypadku rozkładu wykładniczego. Daje się to łatwo wytłumaczyć prawdziwością hipotezy zerowej. W miarę oddalania parametru n od jedynki uzyskujemy wzrost mocy, co jest spowodowane reakcją kształtu funkcji nadwyżki szkody na zmianę tego parametru.

		λ				
	$\tilde{\theta}$	0.5	0.7	1	1.3	1.5
n	0.2	0.823	0.863	0.866	0.893	0.993
	0.5	0.401	0.322	0.314	0.342	0.353
	1	0.027	0.012	0.082	0.003	0.052
	1.5	0.157	0.081	0.138	0.113	0.083
	2	0.342	0.245	0.036	0.613	0.524

Tabela 6.2: Wartości uśrednionej mocy testu dla stuelementowej próby losowej z rozkładu $\mathcal{G}(n, \lambda)$ w zależności od n oraz λ. *Źródło: opracowanie własne.*

Warto zauważyć, że wyniki dla tak krótkich próbek zaburzane są przez użycie metody *bootstrap*. Poniżej prezentujemy tabelę 6.3 z wynikami otrzymanymi przy tych samych założeniach, lecz z użyciem generatora zmiennych losowych z rozkładu gamma w miejscu metody *bootstrap*.

		λ				
	$\tilde{\theta}$	0.5	0.7	1	1.3	1.5
n	0.2	0.753	0.725	0.763	0.722	0.723
	0.5	0.127	0.093	0.125	0.132	0.117
	1	0.053	0.049	0.052	0.047	0.054
	1.5	0.126	0.117	0.113	0.137	0.141
	2	0.355	0.387	0.343	0.324	0.283

Tabela 6.3: Wartości uśrednionej mocy testu (w wersji z użyciem generatora zmiennych losowych) dla stuelementowej próby losowej z rozkładu $\mathcal{G}(n, \lambda)$ w zależności od n oraz λ. *Źródło: opracowanie własne.*

Widzimy tutaj, że moc testu jest podobna w obrębie tej samej wartości n. Otrzymujemy zgodnie z wcześniejszymi uwagami większą moc testu, gdy n jest odległa od jedynki. Nie wydaje się natomiast, żeby parametr λ miał istotny wpływ na moc testu.

6.1.3 Rozkłady Weibulla

Rozkład Weibulla podobnie jak rozkłady gamma jest charakteryzowany przez dwa parametry. Również tutaj rozważymy zatem zależnośc mocy testu od obu z nich. Tabela 6.4 przedstawia uśrednioną moc testu obliczoną z pomocą metody *bootstrap* na podstawie 150 powtórzeń i stuelementowej próbki z rozkładu $\mathcal{W}ei(\tau, \lambda)$.

Ponieważ rozkład wykładniczy jest specjalnym przypadkiem rozkładu Weibulla, znowu możemy obserwować spadek mocy dla $\tau \approx 1$. Im bardziej oddalamy się od niego, tym bardziej moc rośnie - podobnie jak w przypadku rozkładów gamma. Widzimy również dużą moc testu dla $\tau < 1$, czyli rozkładów Weibulla o grubym ogonie. Moc testu jest tam stabilna dla różnych wartości λ, nawet mimo użycia metody *bootstrap*. Nie można tego niestety powiedzieć o lekkoogonowej wersji tego rozkładu - tam moc waha się od 0.041 nawet do 0.793. Jest to jednak znów kwestia związana z użyciem metody *bootstrap*. Generując zmienne losowe prosto z generatora otrzymujemy wyniki widoczne w tabeli 6.5. Znów możemy zaobserwować, że moc testu nie zależy istotnie od parametru skali. To parametr kształtu gra tu kluczową rolę, a moc testu jest zależna przede wszystkim od jego

		λ				
	$\tilde{\theta}$	0.5	0.7	1	1.3	1.5
τ	0.2	0.883	1	1	0.912	1
	0.5	0.903	0.813	0.917	0.767	0.927
	1	0.057	0.317	0.048	0.073	0.011
	1.5	0.182	0	0.107	0.042	0.143
	2	0.153	0.041	0.523	0.433	0.793

Tabela 6.4: Wartości uśrednionej mocy testu dla stuelementowej próby losowej z rozkładu $\mathcal{W}ei(\tau, \lambda)$ w zależności od τ oraz λ. *Źródło: opracowanie własne.*

		λ				
	$\tilde{\theta}$	0.5	0.7	1	1.3	1.5
τ	0.2	0.942	0.943	0.961	0.933	0.947
	0.5	0.363	0.442	0.453	0.383	0.412
	1	0.053	0.047	0.052	0.043	0.033
	1.5	0.103	0.117	0.107	0.156	0.147
	2	0.287	0.28	0.263	0.243	0.267

Tabela 6.5: Wartości uśrednionej mocy testu (w wersji z użyciem generatora zmiennych losowych) dla stuelementowej próby losowej z rozkładu $\mathcal{W}ei(\tau, \lambda)$ w zależności od τ oraz λ. *Źródło: opracowanie własne.*

odległości od jedynki. Zwracamy uwagę na to, że z wyników zaprezentowanych w tabeli 6.5 wygląda na to, że zależność ta nie jest symetryczna. Ruch τ o pewną wartość w dół wydaje się zmieniać moc w inny sposób niż ruch τ o taką samą wartość w przeciwnym kierunku.

6.1.4 Rozkłady lognormalne

Ostatnią grupą rozkładów które rozpatrzymy są rozkłady lognormalne. Ponownie biorąc 150 powtórzeń testu z użyciem metody *bootstrap* otrzymujemy wyniki zaprezentowane dla różnych wartości parametrów μ i σ w tabeli 6.6. Symulacyjnie wyznaczona moc dla

		σ				
	$\tilde{\theta}$	0.1	0.2	0.5	1	2
μ	-0.5	1	1	0.987	0.137	0.997
	0	1	1	0.807	0.11	0.973
	0.5	1	1	0.365	0.893	1
	1	1	1	0.543	0	0.923
	1.5	1	1	0.457	0.0167	1

Tabela 6.6: Wartości uśrednionej mocy testu dla stuelementowej próby losowej z rozkładu $\mathcal{LN}(\mu, \sigma)$ w zależności od μ oraz σ. *Źródło: opracowanie własne.*

rozkładów lognormalnych jest bardzo wysoka. Obserwując wyniki możnaby dojść do wniosku, że jedynie dla pewnych wartości σ test będzie miał kłopot z klasyfikacją ogonu. Jest to uzasadnione stwierdzenie, jeśli pamiętamy, że każdy rozkład lognormalny jest ciężkoogonowy.

Wyniki otrzymane z użyciem generatora zmiennych losowych zamiast metody *bootstrap* przedstawiają natomiast jeszcze wyższą spójność. Widzimy tu, że moc testu zależy przede wszystkim od parametru σ, a parametr μ zmienia bardzo niewiele.

		σ				
	$\tilde{\theta}$	0.1	0.2	0.5	1	2
μ	-0.5	1	1	0.643	0.337	0.957
	0	1	1	0.642	0.328	0.957
	0.5	1	1	0.597	0.267	0.967
	1	1	1	0.577	0.325	0.957
	1.5	1	1	0.664	0.333	0.973

Tabela 6.7: Wartości uśrednionej mocy testu (z wykorzystaniem generatora zmiennych losowych) dla stuelementowej próby losowej z rozkładu $\mathcal{LN}(\mu, \sigma)$ w zależności od μ oraz σ. *Źródło: opracowanie własne.*

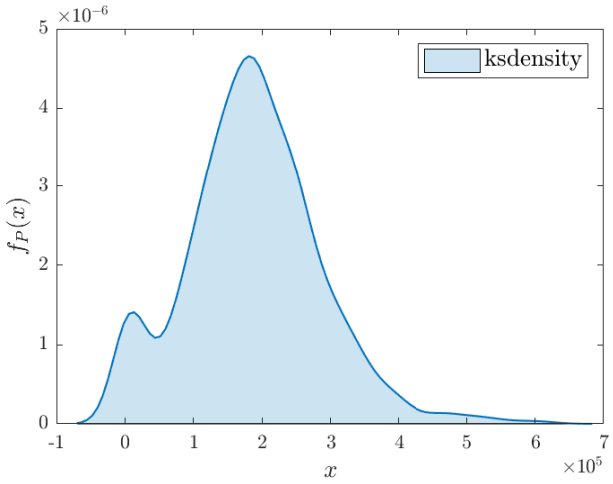
Jak widzimy z testów przeprowadzonych powyżej, w zależności od testowanej próbki algorytm może mieć skrajnie różną moc. Wynika to zarówno z tego że ogony różnych rodzin rozkładów zanikają na różne sposoby, jak i od tego że pracujemy na jednej próbce, co przy skończonej liczbie jej elementów nie gwarantuje, że będzie dobrze odzwierciedlać każdą własność rozkładu z którego pochodzi. Widzimy również jak istotną rolę gra metoda *bootstrap*, która jest konieczna gdy mamy dostępną pojedynczą próbkę, lecz potrafi jednocześnie rozmazać obraz jaki malują przed nami wyniki.

6.2 Wykorzystanie testu do analizy szkodowych danych ubezpieczeniowych

W tym podrozdziale chcemy odejść od komfortu który mieliśmy w poprzednich przykładach. Nie będziemy używać znanych rozkładów zmiennych losowych, co wcześniej pozwalało nam zweryfikować otrzymany wynik z oczekiwaniami. Zderzymy się natomiast z realną sytuacją w której mamy dostępny jedynie wektor danych i brak informacji o rozkładzie któremu podlega.

Wykorzystamy w tym celu dane z sektora ubezpieczeń. Udostępniony został nam fragment bazy pewnego polskiego towarzystwa ubezpieczeniowego dotyczący wypłat świadczeń z tytułu posiadania polis: odpowiedzialności cywilnej (OC), oraz autocasco (AC). Dostępne mamy dwa wektory danych, po jednym dla każdego typu ubezpieczenia, z rekordami na kwoty wypłat dokonane w okresie od 02.01.1998, do 30.12.2000. W sumie uzyskujemy w ten sposób 25869 obserwacji dla OC, oraz 15033 obserwacji dla AC. Dane tego typu można rozpatrywać na wiele sposobów, leczy my skupimy się na dwóch rozkładach powstałych na ich podstawie.

Pierwszym rozkładem który postanowiliśmy rozpatrzyć jest rozkład dziennych sum wypłat z obu ubezpieczeń. W tym celu wypłaty z AC i OC zsumowane zostały w obrębach kolejnych dni, po czym usunięto z wektora obserwacje zerowe, czyli dni bez żadnych wypłat. Powstały w ten sposób wektor danych (nazywany dalej wektorem P) ma długość 833, średnią $\mu_P \approx 187189.6$ oraz odchylenie standardowe $\sigma_P \approx 98463$. Rysunek 6.2 na którym widnieje empiryczna gęstość rozkładu ukazuje nam, że jest ona bimodalna. Ukształtowała się w ten sposób, ponieważ jest mieszanką dwóch próbek o różnych średnich. Z tego powodu



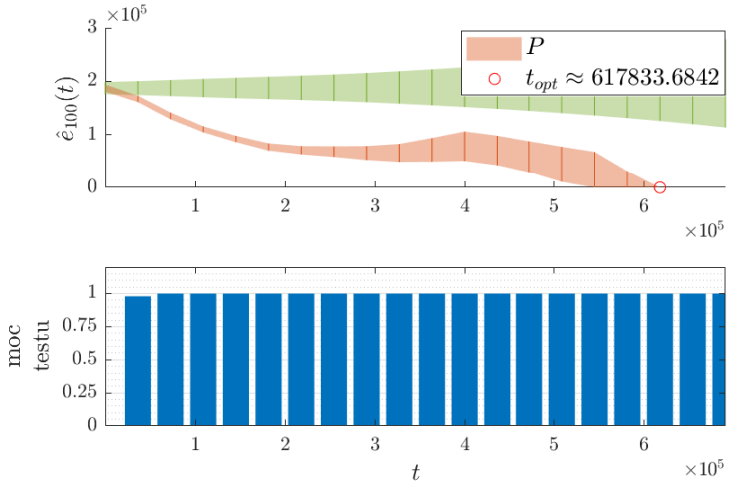
Rysunek 6.2: Empiryczna gęstość $f_P(t)$ rozkładu dziennych sum wypłat z ubezpieczeń OC i AC

nie ma sensu dopasowywanie do niej opisanych wcześniej rozkładów - żaden nie modeluje dobrze dwumodalnych danych.

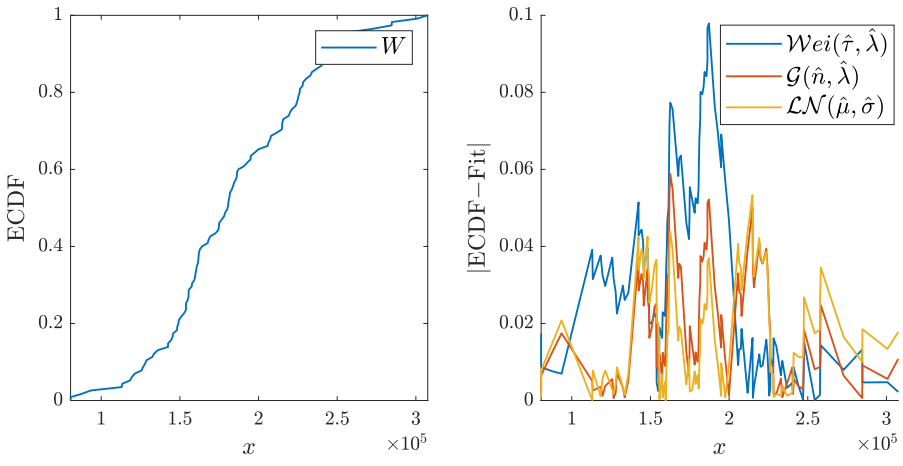
Widzimy również, że ogon próbki zanika dość szybko, co sugeruje nam, że rozkład dziennych sum wypłat z OC i AC może mieć lekki ogon. Poddaliśmy próbkę testowi 5.2. Wynik jego działania widoczny jest na rysunku 6.3. Wcześniejsza uwaga okazała się trafna, co pokazuje algorytm ewidentnie wskazując na lekki ogon. Optymalna wartość parametru t została wybrana jako $t_{opt} \approx 617833.68$, co dla testowanej próbki skutkuje odrzuceniem hipotezy zerowej ze statystyką testową $M = 0$ w lewym obszarze krytycznym oraz p-wartością rzędu 10^{-14}. Z bardzo dużą pewnością można zatem implikować lekki ogon dziennych sum wypłat.

Drugi rozkład który zdecydowaliśmy się rozważyć to rozkład tygodniowych maksimów wypłat z OC. Okres 2 lat jaki obejmuje próbka OC został podsumowany podobnie jak w przypadku poprzedniej próbki w formie dziennych sum. Powstały w ten sposób wektor podzieliliśmy na grupy po 7 kolejnych obserwacji odpowiadających kolejnym raportowanym tygodniom, z których zapisywaliśmy tylko maksymalną w danym tygodniu. Otrzymaliśmy w ten sposób wektor maksimów W o długości 116 obserwacji. Średnia próbkowa wynosi tu $\mu_W \approx 188114.6$, a odchylenie standardowe $\sigma_W \approx 53444$. Na lewym panelu rysunku 6.4 widzimy empiryczna dystrybuantę (*ECDF - Empirical cumulative distribution function*) próbki W. Widzimy na niej, że rozkład jest jednomodalny o dodatnim nośniku - spróbowaliśmy więc dopasować do próbki rozkłady opisane w rozdziale czwartym. Rozkłady dopasowały się bardzo dobrze, co widać na prawym panelu przedstawiającym różnice między ich teoretyczną dystrybuantą a dystrybuantą empiryczną próbki. Najgorzej dopasował się rozkład Weibulla, który zdecydowanie daje największe odchylenia od empirycznego rozkładu. Rozkłady: lognormalny oraz gamma są natomiast lepiej dopasowane i prawie nieodróżnialne od siebie patrząc pod kątem otrzymywanych błędów.

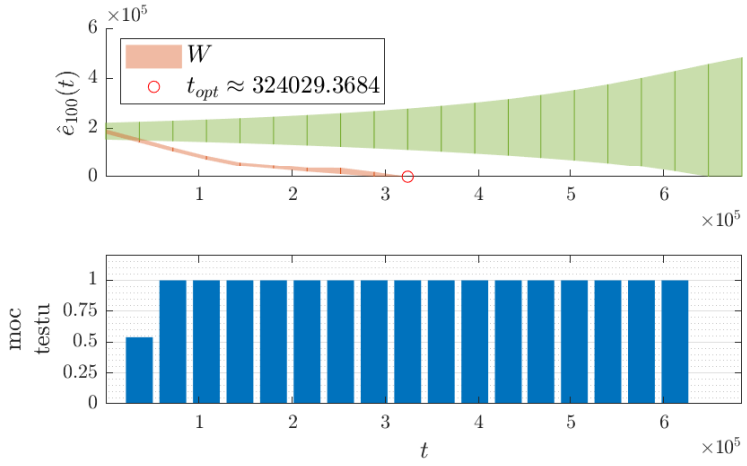
Poddaliśmy następnie próbkę testowi klasyfikującemu ogon. Wynik działania procedury można zobaczyć na rysunku 6.5. Znowu widzimy tutaj ewidentny szybkie przesunięcie empirycznych przedziałów ufności w stronę lekkiego ogonu próbki. Jako optymalne miejsce wykonania testu otrzymaliśmy wartość $t_{opt} \approx 324029.37$, przy której nasza statystyka testowa wpada w lewy obszar krytyczny odpowiadający lekkiemu ogonowi z p-wartością rzędu 10^{-10}. Można zatem z bardzo dużą pewnością podsumować, że tygodniowe maksima



Rysunek 6.3: Algorytm wyboru optymalnego parametru t dla próbki dziennych sum wypłat z AC i OC (próbka P). Górny panel: teoretyczne i empiryczne przedziały ufności estymatora funkcji nadwyżki szkody na poziomie ufności 95%. Dolny panel: symulacyjnie wyznaczona moc testu.



Rysunek 6.4: Lewy panel: empiryczna dystrybuanta tygodniowych maksimów wypłat z OC (próbka W). Prawy panel: odległość empirycznej dystrybuanty od dystrybuant dopasowanych rozkładów



Rysunek 6.5: Algorytm wyboru optymalnego parametru t dla próbki tygodniowych maksimów wypłat z OC (próbka W). Górny panel: teoretyczne i empiryczne przedziały ufności funkcji nadwyżki szkody na poziomie ufności 95%. Dolny panel: symulacyjnie wyznaczona moc testu.

wypłat z ubezpieczenia OC również pochodzą z rozkładu o lekkim ogonie.

Zamykając rozdział o zastosowaniach algorytmu w praktyce nie można nie wspomnieć o istotnym ograniczeniu jaki stanowi moc obliczeniowa komputera. Przeprowadzenie testu zajmuje małą ilość czasu, bez znaczenia jest długa jest próbka. Natomiast wykonywanie graficznej analizy (takiej jaka prezentowana była w pracy) niestety nie jest już tak trywialne przy długich wektorach (tysiąc obserwacji i więcej). Powoduje to symbol Newtona pojawiający się we wzorze na teoretyczny rozkład funkcji nadwyżki szkody, który bardzo szybko eksploduje przyjmując coraz większe wartości. Jeżeli próbka będzie zbyt duża, to możemy po prostu nie być w stanie ich policzyć, co uniemożliwi graficzną inspekcję przedziałów ufności.

Rozdział 7

Podsumowanie

W pracy zaprezentowana została metoda wychwytywania jednej z podstawowych cech rozkładów zmiennych losowych jaką jest grubość ich ogonów. Mimo, że w literaturze istnieje wiele pozycji dotyczących statystycznych metod badania ciężkości ogonu (na czele z *Generalized Extreme Value Theory*, czy *Peaks-Over-Threshold* [7]), to użycie funkcji nadwyżki szkody pozwala spojrzeć na problem w nowy, świeży sposób. Porównując empiryczny rozkład estymatora tej zmiennej losowej dla badanej próbki z jej rozkładem dla próbki z rozkładu wykładniczego jesteśmy w stanie ze sporą pewnością określić rodzaj ogonu, który przejawiają badane obserwacje.

Oprócz prezentacji tej metodologii w pracy wyprowadziliśmy jawny wzór na dystrybuantę estymatora funkcji *MEF* obliczanego na podstawie próbki *iid* z rozkładu $\mathcal{E}xp(\lambda)$, co wprost prowadzi do sposobu nie tylko na klasyfikację ogonu próbki ale również na testowanie zgodności z rozkładem wykładniczym, czy testowanie rozkładu pod kątem ciężkoogonowości lub lekkoogonowości. Zaproponowaliśmy algorytm doboru parametrów testu tak, aby zmaksymalizować jego moc. Zaprezentowaliśmy również działanie testu w rodzinach wybranych rozkładów zmiennych losowych. W pracy zobaczyć można, że w zależności od klasy rozkładów i ich ciężkości ogonu algorytm może osiągać moc od bliskiej 0% do nawet 100%. Pokazaliśmy graficzną analizę działania testu dzięki której często nawet w przypadkach niskiej mocy testu jesteśmy w stanie graficznie określić w stronę którego reżymu skłania się próbka.

Test został też zaaplikowany do danych rzeczywistych - szkodowych danych ubezpieczeniowych. Zbadane zostały rozkłady: dziennej sumy wypłat ubezpieczyciela z tytułu ubezpieczeń AC i OC, oraz tygodniowych maksimów wypłat z ubezpieczenia OC. W obu przypadkach hipoteza zgodności rozkładów z rozkładem wykładniczym została odrzucona. P-wartości statystyk testowych osiągnęły rzędy wielkości odpowiednio 10^{-14} oraz 10^{-10}, co prowadziło do odrzucenia hipotezy na bardzo dużym poziomie istotności. Ponieważ statystyki testowe znalazły się w lewym obszarze krytycznym, test sklasyfikował ogony próbek jako lekkie.

Jak można było zauważyć zaletą proponowanego testu jest prostota jego konstrukcji, zrozumiałość i łatwość implementacji. Do jego wad należy natomiast czas wykonywania, oraz fakt, że dla długich próbek podczas obliczeń zaczynają pojawiać się bardzo duże liczby (rosnące bardzo szybko w związku z użyciem symbolu Newtona) które narażają nas na błędy numeryczne, przeciążenie pamięci i inne niewygodności wynikające z ograniczeń komputerów.

Ponadto pojawia się pytanie na ile poprawne jest używanie testu w przypadku wektora danych które nie są niezależne. Jest to ważne pytanie ze strony praktycznej, ponieważ

zdarzenia ekstremalne, podobnie jak nieszczęścia, chodzą parami *(clustering)* co może być fałszywie odczytywane jako przejaw ciężkiego ogonu.

Praca w żadnym stopniu nie wyczerpuje dyskusji na temat ciężkości ogonu rozkładu. Pojęcie to nadal interpretowane jest na wiele sposobów, a my zgłębiamy zaledwie jedno z możliwych podejść do jednej z możliwych definicji. Jest to zatem wciąż temat otwarty, a co więcej jak najbardziej aktualny i użyteczny w praktyce. Czy kiedyś ktoś postawi kropkę nad „i" ostatecznie rozwiązując kwestię ciężkości ogonu? To pytanie niestety zostaje wciąż bez odpowiedzi i czeka na swój przełomowy moment.

Bibliografia

[1] Borak, S., Hardle, W., Weron, R. (2005). Stable Distributions, *Statistical Tools for Finance and Insurance*. Springer-Verlag, Berlin, str. 21–44.

[2] Burnecki, K., Janczura, J., Weron, R. (2011). Building loss models, Borak, S., Hardle, W., Weron, R. (ed.) *Statistical Tools for Finance and Insurance*. Springer-Verlag, Berlin, str. 293-328.

[3] Burnecki, K., Teuerle, M. (2010). Ruin probability in finite time. *HSC Research Reports HSC/10/04*. Hugo Steinhaus Center, Wrocław.

[4] Cooke, R. (2011). Heavy-tailed distributions: Data, diagnostics, and new developments. *Resources for the Future Discussion Paper No. 11-19*.

[5] Ghosh, S., Resnick, S. (2009). A discussion on mean excess plots. *Stochastic Processes and their Applications 120* , str. 1492-1517.

[6] Guess, F., Proschan, F. (1985). Mean residual life: Theory and applications. *AFOSR Technical Report No. 85-178*. The Florida State University, Tallahassee, Florida.

[7] Jorion, P. (2009). *Financial Risk Manager Handbook, 5th Edition*. Wiley, New Jersey.

[8] Koronacki, J., Mielniczuk, J. (2001). *Statystyka dla studentów kierunków technicznych i przyrodniczych*. Wydawnictwa Naukowo-Techniczne, Warszawa.

[9] Levien, R., Tan, S. (1993). Double pendulum: An experiment in chaos. *American Journal of Physics 61*, str. 1038-1044.

[10] Michailidis, G., Stoev, S. (2012). Extreme value theory: An introduction. *Technometrics 49*, str. 491-492.

[11] Ross, S.(2006). *Simulation*. Elsevier Inc., Amsterdam.

[12] Debo, A. (2010). Probability theory: Stat310/math230. `https://web.stanford.edu/~montanar/TEACHING/Stat310A/lnotes.pdf`, data dostępu: (03-12-2019).

[13] Chang, K. (2008). Edward N. Lorenz, a meteorologist and a father of chaos theory, dies at 90. `https://www.nytimes.com/2008/04/17/us/17lorenz.html`, data dostępu: (05-01-2020).