



ANALIZA DANYCH W UBEZPIECZENIACH

Wanda Ronka-Chmielowiec

Akademia Ekonomiczna Wrocław, Katedra Inwestycji Finansowych i Ubezpieczeń

Ewa Poprawska

Akademia Ekonomiczna Wrocław, Katedra Inwestycji Finansowych i Ubezpieczeń

Wprowadzenie

W teorii aktuarialnej ryzyko ubezpieczeniowe jest sformalizowane i modelowane. W praktyce ubezpieczeniowej do opisu ryzyka ubezpieczeniowego przydatna jest zmienna losowa, która przyjmuje wartości nieujemne. Uzasadnione jest to tym, że następujące zmienne opisujące ryzyko ubezpieczeniowe nieznanne są ubezpieczycielowi i występują losowo:

- ◆ Wartości szkód,
- ◆ Liczby szkód,
- ◆ Momenty czasowe, w których występują szkody lub straty.

Do opisu liczby szkód zastosowanie znajdują zmienne losowe skokowe, które mają skończony lub przeliczalny zbiór wartości pochodzących ze zbioru liczb naturalnych, a do opisu wartości szkód służą zmienne losowe ciągłe, które mogą przybierać dowolne wartości rzeczywiste z pewnego przedziału liczbowego.

Ze zmienną losową związana jest jej funkcja rozkładu prawdopodobieństwa, zwana dystrybuantą, która dobrze charakteryzuje ryzyko ubezpieczeniowe. Stąd dobrymi miarami ryzyka ubezpieczeniowego stosowanymi w praktyce są następujące parametry rozkładu prawdopodobieństwa:

- ◆ Wartość oczekiwana,
- ◆ Odchylenie standardowe,
- ◆ Wariancja,
- ◆ Współczynnik zmienności,
- ◆ Współczynnik skośności.

Jeżeli ryzyko jest mierzalne i w wyniku obserwacji zebrano dostatecznie dużą liczbę danych statystycznych, to wówczas można stosować empiryczne metody modelowania, które wykorzystują statystykę opisową. Wówczas jako miarę ryzyka można wykorzystać parametry pozycyjne, takie jak kwartyle, a zwłaszcza medianę, dominantę, odchylenie ćwiartkowe i inne.



Rozkłady prawdopodobieństwa wykorzystywane do charakteryzacji liczby szkód

W klasycznym ujęciu przyjmuje się, że dobrym modelem opisującym liczbę szkód jest zmienna losowa dyskretna, którą oznaczmy symbolem K . Wówczas zakłada się, że zmienna losowa K przyjmuje wartości $0, 1, 2, \dots$, które oznaczają liczby szkód wygenerowane przez jedno ryzyko lub portfel ryzyk w badanym okresie, z odpowiednimi prawdopodobieństwami p_0, p_1, p_2, \dots .

W praktyce aktuarialnej do charakteryzacji zmiennej losowej K znajdują zastosowanie następujące rozkłady prawdopodobieństwa:

- ◆ Dwumianowy,
- ◆ Poissona,
- ◆ Ujemny dwumianowy,
- ◆ Logarytmiczny.

W przypadku, gdy portfel ubezpieczeń zawiera N jednorodnych rodzajów ryzyka, a prawdopodobieństwo wystąpienia szkody w każdym z nich w ciągu roku jest takie samo i wynosi p , a prawdopodobieństwo nie wystąpienia szkody wynosi $q=1-p$, to zmienna K opisująca liczbę szkód w ciągu roku w tym portfelu jest zmienną losową o rozkładzie dwumianowym gdzie:

$$p^k = P(K = k) = \binom{N}{k} p^k q^{N-k},$$

dla $k=0, 1, 2, \dots, N$.

Przeciętną liczbę przewidywanych szkód w ciągu roku dla powyższego portfela złożonego z N rodzajów ryzyka można oszacować za pomocą wartości oczekiwanej zmiennej losowej K , która dla rozkładu dwumianowego przyjmuje wartość:

$$E(K) = Np.$$

Natomiast wahania rocznej liczby szkód wokół tej przeciętnej ocenia się na podstawie odchylenia standardowego lub wariancji, której wartość określa wzór:

$$\text{Var}(K) = Npq.$$

W praktyce ubezpieczeniowej na ogół występują duże portfele ryzyka, a w ubezpieczeniach majątkowych przeważnie prawdopodobieństwo wystąpienia szkody w pojedynczym ryzyku jest bardzo małe. W takiej sytuacji na podstawie znanych twierdzeń granicznych z rachunku prawdopodobieństwa przyjmuje się, że zmienna losowa K ma rozkład Poissona o postaci:

$$p^k = \frac{\lambda^k}{k!} e^{-\lambda},$$

gdzie $\lambda = Np$, $k=0, 1, 2, \dots$.



W tym przypadku wartość oczekiwana i wariancja zmiennej K określone są następującymi wzorami:

$$E(K)=\lambda, \quad \text{Var}(K)=\lambda.$$

Ponadto w praktyce ubezpieczeniowej mogą pojawić się sytuacje, gdy wariancja liczby szkód przekracza średnią, i wówczas lepszym modelem opisującym zmienną K jest zmienna losowa o rozkładzie ujemnym dwumianowym o następującej postaci

$$p_k = \binom{\alpha + k - 1}{k} q^k p^\alpha$$

gdzie:

$$0 < \alpha < \infty, \quad p \in (0, 1),$$

p – prawdopodobieństwo wystąpienia szkody, $q=1-p$,

p_k – prawdopodobieństwo wystąpienia α -tej kolejnej szkody po k „bezszkodach”.

Dla tego rozkładu podstawowe parametry mają postać:

$$E(K) = \alpha \frac{q}{p}, \quad \text{Var}(K) = \alpha \frac{q}{p^2}.$$

Przykład

Zakład ubezpieczeń w pewnym roku sprzedał 420 polis tego samego typu dla przedsiębiorstw. Zmienna losowa K opisuje liczbę szkód pochodzących z pojedynczej polisy w ciągu roku. Dane liczbowe zestawiono w tabeli 1.

Tabela 1.

Liczba szkód k	0	1	2	3	4	5	6
Liczba polis	126	105	82	60	25	12	10

Oczekiwana liczba szkód pochodzących z jednej polisy wyniesie $E(K)=1,5926$, a wariancja $\text{Var}(K)=3,8725$. Jest to przykład zmiennej, dla której odpowiednim modelem będzie rozkład ujemny dwumianowy. Zgodność rozkładu empirycznego i teoretycznego można sprawdzić za pomocą testu χ^2 .

Innym rozkładem stosowanym do charakteryzacji zmiennej losowej K jest rozkład logarymiczny o następującej postaci:

$$p^k = \frac{p^k}{k |\ln(1-p)|},$$

dla $k=1,2,\dots$, gdzie $p \in (0, 1)$, $q=1-p$.



Podstawowe parametry mają następującą postać:

$$E(K) = \frac{p}{|\ln q|} \frac{1}{q}, \quad \text{Var}(K) = \frac{p}{|\ln q|^2} \frac{1}{q} (|\ln q| - p).$$

Istotną sprawą, którą należy uwzględnić, analizując liczbę szkód lub liczbę wypadków ubezpieczeniowych pochodzących bądź z pojedynczej polisy, bądź z portfela polis, jest fakt, że w praktyce ubezpieczeniowej występuje zmienność intensywności występowania wypadków ubezpieczeniowych, a tym samym szkód. Nawet w tym samym badanym okresie występują lepsze i gorsze okresy dla poszczególnych rodzajów ubezpieczeń, np. ulewne deszcze i gradobicia w okresie wiosenno-letnim, włamania do mieszkań w czasie wakacji, stłuczki samochodowe w okresie jesienno-zimowym i inne. Ponadto zmienną opisującą liczbę szkód należy traktować jako pewien zmieniający się w czasie proces, gdyż spowodowane to może być pewnymi uwarunkowaniami zewnętrznymi, takimi jak nowe metody zapobiegania pożarom, zmiany warunków handlowych, zmiany technologiczne, większa prewencja i inne. Zatem do opisu procesu występowania wypadków ubezpieczeniowych oraz szkód należy brać pod uwagę następujące elementy: trendy, cykle, krótkookresowe wahania, fluktuacje czysto losowe i inne.

Uwzględniając zmienność intensywności występowania wypadków ubezpieczeniowych, wprowadza się zmienną losową mieszającą i ostatecznie zmienna losowa opisująca liczbę szkód ma rozkład prawdopodobieństwa mieszany Poissona lub mieszany rozkład ujemny dwumianowy, zwany rozkładem Poly'ego.

Rozkłady prawdopodobieństwa wykorzystywane do charakteryzacji wartości odszkodowań

Zmienna opisująca wielkości odszkodowań charakteryzuje się następującymi cechami:

- ◆ Wielkości szkód są najczęściej małe i średnie, stąd liczba danych obserwowalnych jest duża,
- ◆ Wielkie odszkodowania występują rzadko, czyli z małymi prawdopodobieństwami, i stąd mało jest danych,
- ◆ Wielkości szkód należy analizować w kontekście czasowym i przestrzennym, gdyż one mogą być zmienną czasu i występować ze zróżnicowanymi wartościami w różnych regionach,
- ◆ W przypadku danych z poprzednich i odległych momentów czasowych należy uwzględnić inflację,
- ◆ Wartości odszkodowań są związane z sumą ubezpieczenia, czyli górną granicą odpowiedzialności ubezpieczyciela lub z decyzjami reasekuracyjnymi, które ustalają udział własny ubezpieczyciela i automatycznie określają górną granicę zbioru obserwacji,



- ◆ Rozkład prawdopodobieństwa wartości tej zmiennej jest przeważnie asymetryczny i w przypadku dużych roszczeń charakteryzuje się tak zwanym wyciągniętym ogonem.
- ◆ Wielkości szkód przyjmują wartości rzeczywiste i dlatego opisują je zmienne losowe ciągle.

Zatem do modelowania wykorzystywana jest zmienna losowa ciągła o pewnej dystrybucji $F(x)$, a przede wszystkim o wartościach $X > 0$, jak również o wartości oczekiwanej $E(X)$ skończonej, gdyż w przeciwnym przypadku takie ryzyko nie jest ubezpieczalne. Wskazane jest również w celu uzyskania ubezpieczalności, aby wariancja także była skończona. Podstawowy problem, jaki tutaj powstaje, to określenie postaci tej dystrybucji lub znalezienie podstawowych parametrów tej dystrybucji. W tym celu stosuje się następujące metody:

- ◆ Konstrukcja empirycznej postaci dystrybucji na podstawie obserwacji z poprzedniego okresu zestawionych w szeregi czasowe,
- ◆ Szukanie analitycznej postaci dystrybucji wyrażonej wzorem i w tym celu wykorzystanie statystycznych testów istotności typu χ^2 lub α -Kołmogorowa lub Kołmogorowa-Smirnowa do zweryfikowania hipotezy o postaci dystrybucji,
- ◆ Obliczanie na podstawie danych empirycznych podstawowych parametrów rozkładu bez szukania postaci dystrybucji.

W zastosowaniach aktuarialnych spotyka się najczęściej następujące typy rozkładów:

- ◆ Gamma,
- ◆ Pareto,
- ◆ Beta,
- ◆ Logarytmiczno-normalny,
- ◆ Normalny.

Pierwszy typ rozkładów to rozkłady gamma o następującej funkcji gęstości:

$$f(x) = \frac{1}{\mu^\gamma \Gamma(\gamma)} x^{\gamma-1} e^{-x/\mu}, \quad x > 0,$$

z dwoma parametrami, gdzie $\mu > 0$, a funkcja gamma Γ zadana jest wzorem:

$$\Gamma(\gamma) = \int_0^{\infty} x^{\gamma-1} e^{-x} dx \quad \text{dla } \gamma > 0.$$

Dla tego rozkładu wartość oczekiwana i wariancja mają postać:

$$E(X) = \mu\gamma, \quad \text{Var}(X) = \mu^2\gamma.$$

Jeśli $\gamma=1$, to otrzymamy szczególny przypadek, zwany rozkładem wykładniczym o następującej postaci funkcji gęstości:

$$f(x) = \frac{1}{\mu} e^{-x/\mu}.$$



Innym rozkładem stosowanym do opisu zmiennej X jest rozkład Pareto.

Dystrybuanta tego rozkładu określona jest wzorem:

$$F(x) = 1 - \frac{1}{(1+x)^\alpha},$$

dla $x \geq 0$ i $0 < \alpha < \infty$.

Podstawowe parametry dla tego rozkładu mają postać:

$$E(X) = \frac{1}{\alpha - 1} \quad \text{dla } \alpha > 1,$$

$$E(X^2) = \frac{2}{(\alpha - 1)(\alpha - 2)} \quad \text{dla } \alpha > 2.$$

W praktyce ubezpieczeniowej stosowany jest również ucięty rozkład Pareto o następującej postaci dystrybuanty:

$$F(x) = \begin{cases} 1 - \frac{1}{(1+x)^\alpha} & \text{dla } 0 \leq x < M < \infty \\ 1 & \text{dla } x \geq M \end{cases}$$

gdzie M nazywa się maksymalną możliwą stratą.

Funkcja gęstości rozkładu Pareto jest określona wzorem:

$$f(x) = \begin{cases} \alpha \frac{1}{(1+x)^\alpha} & \text{dla } 0 \leq x < M < \infty \\ 0 & \text{dla } x \geq M \end{cases}$$

Innym rozkładem służącym do opisu wartości szkody jest rozkład beta. Funkcja gęstości ma wtedy następującą postać:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1},$$

w której występują dwa parametry α i β , spełniające nierówności $0 < \alpha < \infty$, $0 < \beta < \infty$.

Dwa podstawowe parametry tego rozkładu mają następującą postać:

$$E(X) = \frac{\alpha}{\alpha + \beta},$$

$$E(X^2) = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}.$$

Do opisu wartości szkód ubezpieczeniowych jest wykorzystywany również rozkład logarytmiczno-normalny o następującej postaci funkcji gęstości:



$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right),$$

gdzie $0 < y < \infty$ oraz $-\infty < \mu < \infty$, $0 < \sigma < \infty$.

Podstawowe parametry tego rozkładu mają postać:

$$E(X) = e^{\mu + \frac{\sigma^2}{2}}, \text{Var}(X) = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1).$$

W niektórych sytuacjach, gdy portfel ubezpieczeń zawiera bardzo dużo polis i dużo jest danych, to na podstawie centralnego twierdzenia granicznego do opisu wartości wypłaconych odszkodowań lub świadczeń pochodzących z tego portfela może być również wykorzystany rozkład normalny.

Funkcja gęstości ma postać:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad \text{dla } \sigma > 0.$$

Wartość oczekiwana i wariancja dla tego rozkładu wynosi:

$$E(X) = \mu \quad \text{i} \quad \text{Var}(X) = \sigma^2.$$

Przykład

Pewien zakład ubezpieczeń sprzedaje trzy rodzaje ubezpieczeń od kosztów leczenia. Na podstawie danych z poprzednich lat zakład ubezpieczeń przeciętnie w ciągu roku z jednej polisy wypłaca dla pierwszej grupy ubezpieczeń 250 euro z odchyleniem standardowym 120 euro, dla drugiej grupy ubezpieczeń 150 euro z odchyleniem standardowym 80 euro, a dla trzeciej grupy 360 euro z odchyleniem standardowym 240 euro. W bieżącym roku zakład ubezpieczeń sprzedał 340 polis z pierwszej grupy, 650 polis z drugiej grupy oraz 130 polis z trzeciej grupy. Dane liczbowe zestawiono w tabeli 2.

Tabela 2.

Rodzaj ubezpieczenia od kosztów leczenia	1	2	3
Liczba polis	340	650	130
Oczekiwane świadczenie z pojedynczej polisy	250	150	360
Odchylenie standardowe dla pojedynczej polisy	120	80	240

Z oznacza całkowitą wartość wypłaconych świadczeń z portfela.

$$E(Z) = 340 \cdot 250 + 650 \cdot 150 + 130 \cdot 360 = 229300,$$

$$\text{Var}(Z) = 340 \cdot 14400 + 650 \cdot 6400 + 130 \cdot 57600 = 1654400,$$

czyli $\sigma = 4067,43$.



Zakład ubezpieczeń chce mieć informację, jakie jest prawdopodobieństwo, że całkowita kwota świadczeń z tytułu ubezpieczeń od kosztów leczenia w danym roku przekroczy 240 000 euro.

Aby odpowiedzieć na to pytanie, korzystamy z centralnego twierdzenia granicznego, co zapisujemy wzorem:

$$P(Z > 240\,000) = 1 - P(Z \leq 240\,000) = 1 - F_Z(240\,000),$$

gdzie

$$F_Z(x) = \Phi\left(\frac{x - 229300}{4067,43}\right)$$

oraz Φ jest dystrybuantą rozkładu normalnego standaryzowanego. Korzystając z tablic rozkładu normalnego, otrzymamy:

$$1 - F_Z(240000) = 1 - \Phi(2,63) = 0,0043.$$

Do modelowania wartości szkód pochodzących ze zdarzeń katastrofalnych wykorzystywane są rozkłady wartości ekstremalnych. Podamy tutaj przykładowo trzy najczęściej stosowane rozkłady:

- ◆ rozkład Frecheta o następującej dystrybuancie:

$$F(x) = \begin{cases} 0 & \text{dla } x \leq 0 \\ \exp(-x^{-\alpha}) & \text{dla } x > 0 \end{cases}$$

gdzie parametr $\alpha > 0$,

- ◆ rozkład Weibulla o następującej dystrybuancie:

$$F(x) = \begin{cases} \exp(-(-x)^{-\alpha}) & \text{dla } x \leq 0 \\ 1 & \text{dla } x > 0 \end{cases},$$

gdzie parametr $\alpha > 0$,

- ◆ rozkład Gumbela o następującej dystrybuancie

$$F(x) = \exp(-\exp(-x)).$$

W przypadku ryzyka katastrofalnego pojawiają się następujące problemy związane z jego oceną i pomiarem:

- ◆ nieporównywalność danych opisujących zdarzenia katastrofalne z powodu zmieniającego się otoczenia, zmieniającej się wartości ubezpieczonego majątku, przyrostu lub spadku populacji ludzi zamieszkujących zagrożony teren, zmian klimatycznych, zmieniającej się technologii, występującej inflacji i innych,
- ◆ mała liczba danych o zdarzeniach katastrofalnych, które występują rzadko, i stąd zbyt krótkie historyczne szeregi czasowe, które nie mogą dać pełnej informacji o danym zjawisku katastrofalnym,



- ◆ wykorzystywanie danych z innych dziedzin, szczególnie danych sejsmicznych, meteorologicznych, ekonomicznych i innych do modelowania zdarzeń katastrofalnych,
- ◆ wykorzystanie metod symulacyjnych do oceny rozkładów prawdopodobieństw i ich parametrów zmiennych losowych opisujących zdarzenia katastrofalne.

Szukanie analitycznej postaci dystrybuanty - przykład

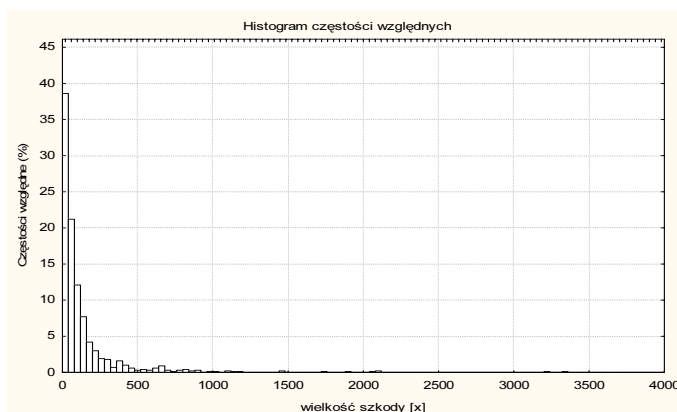
Przykład doboru rozkładu wysokości szkody oparty jest na danych umownych (1000 obserwacji), wygenerowanych z rozkładu Pareto, w którym parametry nie są stałymi, lecz również są zmiennymi losowymi (o rozkładzie Gamma). Wprowadzenie zmiennych mieszających miało na celu uzyskanie danych, które będą symulować zachowanie portfela szkód niejednorodnych, związanych np. z występowaniem zjawisk katastrofalnych.

Do tak przygotowanych danych będą dopasowywane rozkłady, które najczęściej stosowane są do modelowania wysokości odszkodowań. Są to rozkłady: wykładniczy, gamma, logarytmiczno-normalny, Pareto, Weibulla, Burra. W pakiecie *STATISTICA* pierwsze trzy z wymienionych znajdują się wśród rozkładów, jakie można dobierać do danych. Rozkład Weibulla znaleźć można natomiast w *Analizie przeżycia*. Dobór pozostałych możliwy jest przy wykorzystaniu *Estymacji nieliniowej*, gdzie możliwe jest dopasowywanie dowolnej funkcji wprowadzonej przez użytkownika.

Kształt histogramu częstości analizowanych danych (rysunek 1.), przede wszystkim występowanie obserwacji bardzo wysokich w porównaniu z wartością oczekiwaną i wariancją, może wskazywać, że konieczne będzie dopasowanie rozkładu charakteryzującego się tzw. „ciężkim ogonem”.

Parametry rozkładu empirycznego są następujące:

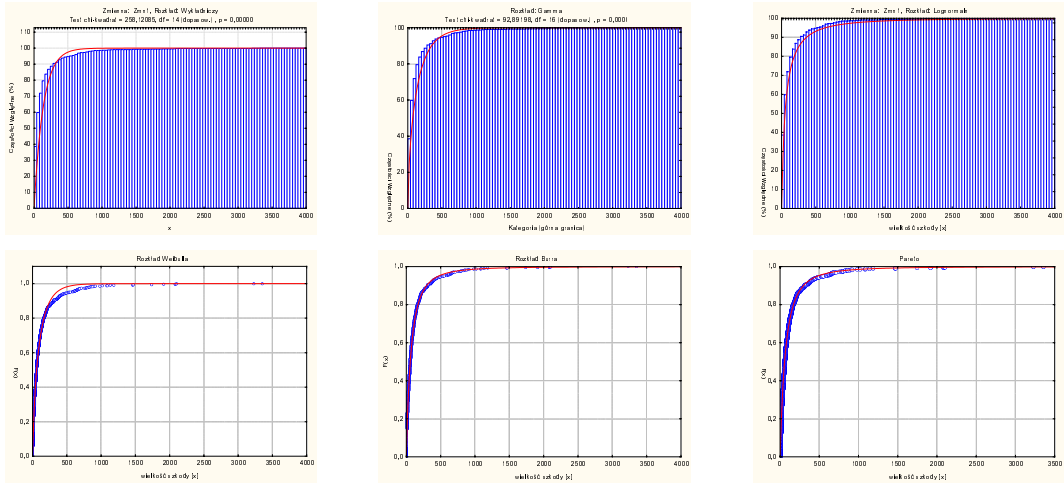
- ◆ średnia obserwowana: 135,37;
- ◆ wariancja obserwowana: 69718,81.



Rys. 1. Histogram częstości względnych

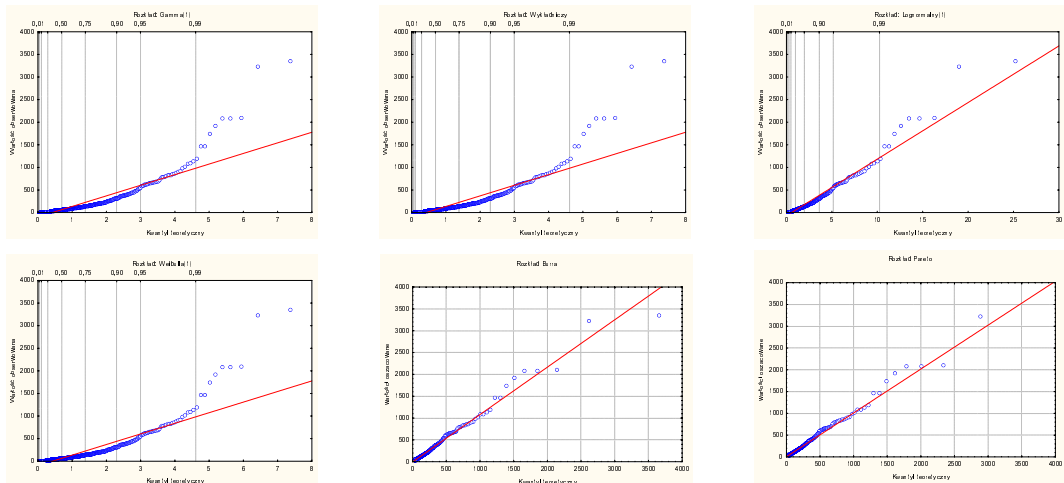


Natomiast poniższe wykresy, przedstawiające porównanie dystrybuanty empirycznej z dystrybuantami sześciu analizowanych rozkładów, pozwalają na wstępny wybór właściwego rozkładu. Zarówno rozkład wykładniczy, jak i gamma oraz Weibulla niedoszacowują prawdopodobieństwa wystąpienia szkód wysokich, czyli tych, które w dużym stopniu wpływają na wysokość wartości oczekiwanej, która z kolei stanowi podstawę szacowania składki ubezpieczeniowej. Natomiast rozkład logarytmiczno-normalny przeszacowuje te prawdopodobieństwa, co skutkuje zawyżeniem składki ubezpieczeniowej. Z tego względu warto przeanalizować pozostałe dwa rozkłady: Burra i Pareto (parametry tych rozkładów zostały dobrane metodą największej wiarygodności przy użyciu *Estymacji nieliniowej*).



Rys. 2. Porównanie dystrybuanty empirycznej z dystrybuantami dobranej rozkładów

Aby dokonać wyboru modelu, który najlepiej opisuje dane, warto przeanalizować wykresy kwantyl-kwantyl.



Rys. 3. Wykresy kwantyl-kwantyl dla dopasowywanych rozkładów

Wykresy kwantyl-kwantyl pozwalają stwierdzić, dla jakich wielkości szkód rozkłady teoretyczne odbiegają od danych, a dla jakich dobrze je opisują. Dla rozkładów dobrze opisujących dane, punkty na wykresie powinny układać się wzdłuż prostej. Dla rozkładów takich jak wykładniczy, gamma czy Weibulla, które niedoszacowują prawdopodobieństwa wystąpienia wysokich szkód, punkty odpowiadające kwantylom empirycznym odchylają się w górę.

Aby dokonać ostatecznego wyboru rozkładu, trzeba przetestować dopasowanie rozkładów do danych. W tym wypadku posłużono się statystyką λ -Kolmogorowa.

Tabela 3.

gamma		Weibulla		wykładniczy		Pareto		logarytmiczno-normalny		Burra	
α	1,001	θ	102,7	θ	105,4	α	6,197	μ	4,082	α	2,093
θ	105,3	τ	0,95			θ	543,8	σ	1,158	θ	133,9
										γ	1,296
λ	2,58	2,15	2,57	2,19	1,9	0,44					

Źródło: obliczenia własne

Na poziomie istotności równym 0,05 tylko dla rozkładu Burra nie przyjmuje ona wartości ze zbioru krytycznego, co ostatecznie pozwala rozstrzygnąć, który z analizowanych modeli najlepiej opisuje dane.

Wyznaczanie składki ubezpieczeniowej metodą największej wiarygodności

Wyjściową zasadą stosowaną przy kalkulacji składki ubezpieczeniowej jest zasada równoważności składki i wypłaconych w przyszłości odszkodowań i świadczeń, zwana zasadą czystego ryzyka. Składka ubezpieczeniowa netto oznaczona symbolem P powinna spełniać następującą równość:

$$P = E(X).$$

W praktyce ubezpieczeniowej zasada ta jest traktowana jako podstawowa i jest na ogół modyfikowana różnymi sposobami zapewniającymi równowagę finansową operacji ubezpieczeniowych w portfelach, które poprawiają bezpieczeństwo finansowe.

Zatem dla zakładu ubezpieczeń istotną sprawą jest określenie składki na odpowiednim poziomie, który będzie zaakceptowany przez obie strony kontraktu ubezpieczeniowego, czyli ubezpieczającego i ubezpieczyciela. Ubezpieczyciel wysokość przyszłej składki określa na podstawie przeszłych obserwacji, na ogół dla całej homogenicznej grupy ryzyka. Zauważyć jednak należy, że po pierwsze grupa może być zróżnicowana, zatem należałoby brać obserwacje dla poszczególnych polis, a po drugie poszczególni ubezpieczający mogą



mieć również własne obserwacje dotyczące ubezpieczonego ryzyka, które można uwzględnić. W takich sytuacjach znajduje zastosowanie teoria wiarygodności.

Powstaje tutaj podstawowe pytanie: jaką dać wiarygodność obserwacjom dotyczącym j-tego ryzyka, a jaką obserwacjom przeszłym pochodzącym z całego portfela.

Aby określić składkę netto w kontrakcie, należy obliczyć następującą wielkość:

$$z_j E(X_j) + (1-z_j)E(X),$$

gdzie X_j opisuje ryzyko w postaci odszkodowania pochodzące z j-tej polisy lub z pewnej j-tej grupy jednakowych polis o takich samych warunkach kontraktu, a X opisuje ryzyko wygenerowane przez cały portfel polis. Natomiast z_j jest współczynnikiem wiarygodności spełniającym następującą nierówność $0 \leq z_j \leq 1$.

Podstawowym problemem, jaki trzeba rozwiązać w przypadku stosowania tego typu podejścia, jest wyznaczenie współczynnika wiarygodności z_j . Współczynnik ten powinien spełniać następujące warunki:

- ♦ im więcej jest dostępnych obserwacji dotyczących danej podgrupy ryzyk, tym większa jest waga tych obserwacji - wyższy współczynnik wiarygodności;
- ♦ im większe są różnice pomiędzy poszczególnymi podgrupami ryzyk, tym większa powinna być waga obserwacji dotyczących tych grup; im bardziej jednorodny jest portfel polis, tym mniej zróżnicowane powinny być składki, w większym stopniu oparte na oczekiwanej wysokości szkody dla całego portfela;
- ♦ im większe są wahania wysokości szkód w obrębie danej podgrupy w czasie, tym większa powinna być waga przypisywana oczekiwanej wysokości szkody dla całego portfela (obserwacje pochodzące wyłącznie z określonej podgrupy polis stają się mało użyteczne dla kalkulacji składki).

Jednym z najprostszych modeli, jakie mają zastosowanie do rozwiązania tego zagadnienia jest zrównoważony model Buhlmana.

W tym modelu zakłada się, że portfel polis ubezpieczeniowym można podzielić na J podgrup, z których każda zawiera jednakową liczbę polis (można również przyjąć, że każda z podgrup składa się tylko z jednej polisy). Ponadto dostępna jest taka sama liczba obserwacji dotyczących szkód występujących w każdej z grup - dla każdej z podgrup dostępne są dane dotyczące T okresów. Czyli X_{jt} oznacza wysokość szkody powstałej w j-tej podgrupie w okresie t . Dla grup składających się z n polis X_{jt} może być interpretowane jako suma szkód powstałych w okresie t w grupie j lub też ich średnia.

Każdą z obserwacji można rozdzielić na sumę trzech niezależnych elementów:

$$X_{jt} = m + \Xi_j + \Xi_{jt}$$

gdzie:

m - średnia wysokość szkody dla całego portfela polis ubezpieczeniowych;

Ξ_j - odchylenie od tej średniej, które charakteryzuje polisy należące do j-tej klasy;



Ξ_{jt} - odchylenie od średniej szkody dla danej klasy, które jest związane z czasem (charakteryzuje „dobre” i „złe” okresy posiadaczy polis ubezpieczeniowych).

Zakłada się, że Ξ_j oraz Ξ_{jt} są niezależnymi zmiennymi losowymi, takimi że

$$E(\Xi_j) = E(\Xi_{jt}) = 0, \quad Var(\Xi_j) = \alpha, \quad Var(\Xi_{jt}) = s^2$$

czyli α charakteryzuje różnice pomiędzy poszczególnymi polisami lub grupami polis.

Rozkład zmiennej losowej Ξ_j opisuje strukturę ryzyka w portfelu. Podobnie rozkład zmiennej losowej Ξ_{jt} opisuje zmiany ryzyka w czasie, stąd parametry α i s^2 mogą być interpretowane jako parametry ryzyka.

Można wykazać, że najlepsza w sensie minimalnego średniego błędu kwadratowego

$$E\left[\left\{X_{j,T+1} - g_{11}X_{11} - \dots - g_{JT}X_{JT}\right\}^2\right]$$

prognoza wysokości szkody dla grupy j w okresie $T+1$, która jest kombinacją liniową obserwacji szkód, jest równa składce wiarygodnej:

$$z\bar{X}_j + (1-z)\bar{X}$$

gdzie

$$z = \frac{\alpha T}{\alpha T + s^2}.$$

Współczynnik wiarygodności dla portfeli polis ubezpieczeniowych jest jednakowy dla wszystkich wyróżnionych podgrup, czyli waga indywidualnego doświadczenia dla każdej grupy jest taka sama.

Ponadto:

$$\bar{X} = \frac{1}{JT} \sum_{j=1}^J \sum_{t=1}^T X_{jt}$$

oraz

$$\bar{X}_j = \frac{1}{T} \sum_{t=1}^T X_{jt}$$

Wartość współczynnika wiarygodności z niezbędnego do kalkulacji składki zależna jest od wartości parametrów ryzyka, charakteryzujących portfel polis ubezpieczeniowych, oraz liczby dostępnych obserwacji szkód dotyczących danej podgrupy. Zatem estymator współczynnika wiarygodności powinien opierać się na wcześniejszej estymacji tych parametrów.

Kalkulacja współczynnika z może być dokonana w oparciu o następujące statystyki:

♦ MSB (mean-square-between):

$$MSB = \frac{\sum_{j=1}^J T(\bar{X}_j - \bar{X})^2}{J - 1}$$

◆ MSW (mean-square-within):

$$MSX = \frac{\sum_{j=1}^J \sum_{t=1}^T (X_{jt} - \bar{X}_j)^2}{J(T-1)}$$

Statystyka określona jako $F = MSB / MSW$ ma wartość oczekiwaną równą $\alpha T + s^2$, natomiast statystyka MSW ma wartość oczekiwaną równą s^2 . Stąd statystyka równa $1 - 1/F$ może być estymatorem współczynnika wiarygodności z .

Zatem kalkulacja składki wiarygodnej oparta jest na stosunkowo prostych obliczeniach – najprostszych statystykach, takich jak wartość oczekiwana zmiennych X_j , czy wariancja obliczone dla danych.

Poniżej przedstawione są wyniki obliczeń wiarygodnej składki dla niewielkiego portfela polis ubezpieczeń domów. Polisy zostały podzielone na 12 podgrup ze względu na dwie kategorie: wiek budynku (podział na 3 kategorie: nowe - budynki młodsze niż 10 lat, stare - powyżej 30 lat, pozostałe) oraz położenie (podział na 4 kategorie). Każda z wyróżnionych podgrup zawiera jednakową liczbę polis ubezpieczeniowych. W tabeli przedstawione są dane dotyczące średniej wysokości szkody w każdej z podgrup. Dostępne dane obejmują okres 10 lat.

Tabela 4. Średnie wysokości szkód w podgrupach polis.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
rok 1	98,88	85,42	94,55	94,13	94,45	101,59	85,41	98,71	96,67	103,22	114,74	89,20
rok 2	109,83	82,27	97,24	100,48	97,89	107,87	99,65	90,62	101,12	102,73	106,28	100,46
rok 3	118,11	96,89	109,48	85,36	108,44	97,51	96,31	115,36	125,93	95,68	97,67	106,70
rok 4	114,24	99,88	96,43	99,51	97,46	102,93	115,42	102,75	86,52	91,20	100,93	80,61
rok 5	109,19	85,95	103,81	98,71	90,46	108,06	97,68	111,12	95,74	86,51	117,91	96,45
rok 6	111,42	85,42	90,92	109,49	91,33	104,80	87,44	101,06	105,41	98,76	100,01	106,77
rok 7	109,02	105,99	111,34	86,48	111,36	98,93	110,69	93,79	106,66	92,60	100,93	98,00
rok 8	104,29	94,29	90,43	108,92	98,22	110,75	100,86	94,06	90,34	90,73	115,33	102,13
rok 9	99,89	105,86	99,39	99,10	90,08	100,96	100,09	105,80	89,17	104,02	100,63	102,34
rok 10	97,01	99,29	93,09	98,51	100,41	97,87	87,33	117,66	103,96	106,42	117,49	109,14

Źródło: obliczenia własne

Dla analizowanego przykładu wartości wymienionych wyżej statystyk są następujące:

$$MSB = 161,43 \quad MSW = 68,55 \quad F = 2,3549$$

Wreszcie wartość współczynnika wiarygodności wynosi $z = 0,575353$. Współczynnik ten jest jednakowy dla wszystkich wyróżnionych podgrup. Wysokość składki wiarygodnej oraz niezbędne obliczenia przedstawione są w tabeli. Średnia szkoda w portfelu wynosi 100,34.

Tabela 5. Obliczenia wiarygodnej składki

Grupa polis	Średnia szkoda w grupie	Wariancja szkód w grupie	Składka
1	107,19	48,4069	104,28
2	94,13	78,3202	96,76
3	98,67	54,1843	99,38
4	98,07	63,2249	99,03
5	98,01	52,2475	98,99
6	103,13	21,2774	101,94
7	98,09	96,2792	99,04
8	103,09	86,8652	101,92
9	100,15	131,5773	100,23
10	97,19	46,2489	98,53
11	107,19	67,6326	104,28
12	99,18	76,3301	99,67

Źródło: obliczenia własne

Zastosowanie nawet bardzo prostych obliczeń pozwala na odpowiednie zróżnicowanie wysokości składki ubezpieczeniowej dla tego typu niejednorodnych polis ubezpieczeniowych. Oczywiście możliwe jest również zastosowanie metod opartych na analizie wariancji ANOVA czy też regresji, bądź wreszcie uogólnionych modeli liniowych (GLM).

Literatura

1. Ronka-Chmielowiec W. (2003): *Modelowanie ryzyka w ubezpieczeniach. Wybrane zagadnienia*. Wyd. AE we Wrocławiu.
2. Bailey A. (1950): *Credibility procedures*, Proceedings of the Casualty Actuarial Society, XXXVII, 7-23, 94-115.
3. Bühlmann H. (1967): *Experience rating and credibility*, ASTIN Bulletin 4, 199-207.
4. Daykin C.D., Pentikäinen T., Pesonen M. (1996): *Practical Risk Theory for Actuaries*, Chapman & Hall, London.
5. Kaas R., Goovaerts M., Dhaene J., Denuit M. (2001): *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, Boston.
6. Klugman S., Panjer H.H., Willmot G.E. (1998): *Loss models: From Data to Decisions*, John Wiley & Sons, New York.
7. Nelder J.A., Verrall R.J. (1997): *Credibility theory and generalised linear models*, ASTIN Bulletin 27 No 1, 71-82.