НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Факультет прикладної математики

Кафедра прикладної математики

«На правах рукопису»
УДК 517.977

«До захисту допущено»

Завідувач кафедри
_____ О. Р. Чертов
(підпис)

«___» _____2015 р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності 8.04030101 «Прикладна математика»

на тему: Оцінка страхових затрат на основі копула-моделей

Виконав: студент 2 курсу, групи КМ-31М

Савін Євген Вадимович _____ (підпис)

Науковий керівник — доцент, канд. техн. наук, доцент Олефір О. С. _____ (підпис)

Консультант із нормоконтролю — старший викладач Мальчиков В. В. _____ (підпис)

Рецензент — професор, д-р техн. наук, проф. Симоненко В. П. _____ (підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2015 року

Національний технічний університет України
«Київський політехнічний інститут»

Факультет прикладної математики

Кафедра прикладної математики

Рівень вищої освіти – другий (магістерський)

Спеціальність 8.04030101 «Прикладна математика»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ О. Р. Чертов
(підпис)
«___» _____2015 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Савіну Євгену Вадимовичу

1. Тема дисертації: «Оцінка страхових затрат на основі копула-моделей», науковий керівник дисертації Олефір Олександр Степанович, канд. техн. наук, доцент,
затверджені наказом по університету від «20» березня 2015 року № 785-С.

2. Термін подання студентом дисертації: «18» червня 2015 р.

3. Об'єкт дослідження: історичні дані збитків нанесених пожежами застрахованим будівлям, майну і матеріалам, що були в цих будівлях та збитків через припинення фінансової діяльності внаслідок пожежі.

4. Предмет дослідження: метод та спроби реалізації покращення розрахунків вартостей страхових полісів за допомогою моделей, що використовують копули.

5. Перелік завдань, які потрібно розробити:

– проаналізувати існуючі підходи моделювання випадкових величин залежностей у страхуванні ;

– обґрунтувати методи та засоби вирішення поставленої задачі ;

– розробити алгоритм моделювання копул з урахуванням наявності важких хвостів у маргінальних розподілів на основі зафіксованих страхових випадках пожеж у будівлях ;

– розробити програмні засоби для реалізації алгоритму та отримати копула-модель збитків при пожежах в будівлях ;

– протестувати якість та асимптотичні властивості моделі ;

– навести приклад розрахунку чистих премій з використанням копула моделей.

6. Орієнтовний перелік ілюстративного матеріалу:

– таблиці класифікації методів знаходження розподілів випадкових величин ;

– огляд існуючих методів перевірки статистичних гіпотез ;

– огляд основних сімейств копул ;

– графіки порівняння функцій розподілів ;

– графіки порівняння копул ;

– результати перевірки якості отриманої моделі.

7. Орієнтовний перелік публікацій:

– Міжнародна наукова конференція імені Т.А. Таран «Інтелектуальний аналіз інформації – ИАИ-15»;

– 17-та міжнародна конференція «Системний аналіз та інформаційні технології – SAIT 2015».

8. Дата видачі завдання «25» жовтня 2013 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Вибір напряму дослідження та узгодження тематики МД з керівником	15 вересня–30 жовтня 2013	
2	Огляд існуючих методів та засобів моделювання залежностей у страхуванні	30 жовтня 2013–15 лютого 2014	
3	Дослідження розподілів обраних випадкових величин	15 лютого–1 вересня 2014	
4	Побудова та покращення копула-моделей на основі обраних величин	1 вересня 2014–1 березня 2015	
5	Розробка та покращення програмних засобів	1 березня–1 травня 2015	
6	Оформлення текстової і графічної частини МД	1 травня–1 червня 2015	
7	Попередній захист МД	1 червня–15 червня 2015	

Студент _____ Є. В. Савін
(підпис)

Науковий керівник дисертації _____ О. С. Олефір
(підпис)

РЕФЕРАТ

Актуальність проблеми. Вимоги моделювання страхових компаній ростуть з кожним роком. Впровадження більш гнучких типів контрактів для захоплення ринку викликає необхідність у більш чіткому моделюванні залежностей між випадковими величинами. Коефіцієнт лінійної кореляції не достатньо відповідає вимогам моделювання. Саме тому, розробка копула-моделей, що можуть пояснювати нелінійні типи залежностей є дуже актуальною в страховому світі.

Об'єкт дослідження – історичні дані данської страхової компанії, що фіксують збитки нанесені пожежами застрахованим будівлям та іншим фінансовим складовим, що знаходились в цих будівлях.

Предмет дослідження – ефективні копула-моделі, побудовані на основі реальних статистичних даних.

Мета роботи – розробка алгоритму побудови ефективних копула моделей з урахуванням можливості наявності важких хвостів в законах розподілу випадкових величин та в застосуванні цих моделей при розрахунку страхових тарифів.

Методи дослідження – використовуються методи параметричної оцінки, зокрема метод найбільшої правдоподібності, непараметричний тест Колмогорова-Смирнова. Програмна реалізація методів виконана в безкоштовному статистичному середовищі R.

Наукова новизна роботи полягає в тому, що:

1. Розроблено принципово новий підхід побудови копула моделей на модифікованих маргінальних розподілах.
2. Вперше запропонована комбінація теореми Пікандса і копула моделей.
3. Розроблено модель, що пояснює збитки нанесені при пожежі будівлі та іншим її фінансовим складовим.

Практична цінність отриманих в роботі результатів полягає в тому, що розроблений алгоритм, дозволяє швидко побудувати точну та якісну модель залежних страхових збитків, на основі якої можна розраховувати страхові тарифи.

Апробація роботи. Основні положення і результати проміжних досліджень були представлені на міжнародній науковій конференції імені Т.А. Таран «Інтелектуальний аналіз інформації – ИАИ-15» та на 17-ій міжнародній конференції «Системний аналіз та інформаційні технології – SAIT 2015».

Структура та обсяг роботи. Магістерська дисертація складається зі вступу, п'яти розділів, висновків та додатків.

У вступі стисло описано загальну характеристику роботи, подано оцінку сучасного стану досліджуваної проблеми, обґрунтовано актуальність дослідження, сформульовано мету і завдання дослідження, наукову новизну отриманих результатів і практичну цінність роботи.

У першому розділі розглянуті основні статистичні методи та підходи для визначення розподілів випадкових величин, побудови копула-моделей та моделювання важких хвостів і екстремальних значень. Розглянуті вхідні дані, що будуть використані для моделювання.

У другому розділі підібрані закони розподілів та оцінені параметри для вхідних випадкових величин, підібрана оптимальна копула модель на основі коефіцієнту Кендала.

У третьому розділі проведено аналіз хвостів отриманої моделі, застосовано теорему Пікандса для покращення маргінальних розподілів та побудована покращена модель.

У четвертому розділі проведений аналіз отриманих результатів.

У п'ятому розділі проведено аналіз використаних програмних засобів та розроблених функцій.

У висновках викладено найбільш значущі наукові та практичні результати проведеного наукового дослідження та програмної реалізації,

обґрунтовано достовірність отриманих результатів.

Загальний обсяг роботи становить 89 сторінок, основний зміст викладено на 59 сторінках. Робота містить 2 додатки, список використаних літературних джерел з 22 найменувань, 8 рисунків та 12 таблиці.

Ключові слова: копула-модель, страхування, актуарні науки, хвости розподілів, залежні випадкові величини.

ABSTRACT

Background. Requirements in modeling of insurance companies are growing every year. The introduction of more flexible types of contracts to capture market necessitates a more precise modeling of dependencies between random variables. The coefficient of linear correlation is not sufficiently meet the requirements of the modeling. Therefore, the development of copula models that can explain the types of nonlinear dependence is very important in the insurance world.

The object of research - historical data of Danish insurance company, fixing damage caused by fire to insured buildings and other financial components that were in these buildings.

Subject of research - effective copula models are based on real statistics.

Purpose of research - development of efficient algorithm for building copula models with the possibility of the presence of heavy tails in the laws of distribution of random variables and applying these models when calculating insurance rates.

Methods of research - parametric estimation methods are used, including the most likelihood method, nonparametric Kolmogorov-Smirnov test. Software implementation techniques performed in free statistical environment R.

Scientific novelty lies in the fact that:

1. A new approach to building copula models with modified marginal distributions.
2. The first time the combination of Pikands theorem and copula models.
3. The model that explains the damage caused by fire to buildings and its other financial components.

The practical value obtained in the results is that the developed algorithm can quickly build accurate models of dependences in insurance losses on which we can calculate insurance rates.

Testing of work. The main provisions of interim studies and the results were presented at the international conference named T.A. Taran "Intelligent analysis of information - IAI-15" and the 17-th International Conference "System Analysis and Information Technology - SAIT 2015".

The structure and scope of work. Master's thesis consists of introduction, five chapters, conclusions and applications.

The introduction briefly describes the general characteristics of the work evaluates the current state of research problem, the urgency of research, formulated goals and research objectives, scientific novelty of the results and practical value of the work.

The first section describes the main statistical methods and approaches to determine the distributions of random variables, copula building models and modeling of heavy tails and extreme values. Considered the input data used for modeling.

The second section describes chosen distribution law and assessed parameters for input random variables, the optimum copula model is chosen on the base of Kendall's coefficient.

In the third section provides analyzes the tails of the resulting model, Pikands theorem is applied to improve marginal distributions and built an upgraded model.

The fourth section provides analysis of the results.

The fifth section analyzes the used software and developed functions.

The conclusions set out the most important scientific and practical results of scientific research and program implementation, proved the reliability of the results.

The total amount of work is 89 pages, the essence contained of 59 pages. Work includes 2 applications, the list of used literature of 22 items, 8 figures and 12 tables.

Keywords: copula model, insurance and actuarial science, tails distributions dependent random variables.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ І ПОЗНАЧЕНЬ

ВВ – Випадкова величина

ММП – Метод максимальної правдоподібності

КС-тест – Тест Колмогорова-Смирнова

ВСТУП

Моделювання випадкових процесів лежить в основі страхового бізнесу. Весь механізм страхування та оцінки ризиків базується на статистичних моделях. Основною ціллю передових актуарних розрахунків являється розробка та калібрування таких моделей, що найкраще повторюють реальні події. Саме тому при створенні повної та максимально точної моделі необхідно враховувати залежності, що існують між випадковими величинами.

Одним з найпоширеніших способів, яким описували та продовжують описувати залежності випадкового типу, являється коефіцієнт лінійної кореляції. Проте гнучкість та точність таких моделей дуже слабка, адже залежність між випадковими величинами в більшості випадків не є лінійною.

Для моделювання складних типів залежностей використовують копули. Вони широко застосовані для оцінки фінансових ризиків і в страховому аналізі – наприклад для ціноутворення забезпечених боргових зобов'язань. Також копули застосовуються до інших страхових задач як гнучкий інструмент. Нещодавно, копули були успішно використані для формування бази даних для аналізу надійності автострадних мостів і для різноманітних моделювань з багатьма змінними в цивільному, механічному та шельфо-видобувному машинобудуванні [1].

1 МОДЕЛЮВАННЯ ЗАЛЕЖНОСТЕЙ В СТРАХУВАННІ

Моделювання залежностей випадкових величин базуються на статистичному аналізі зібраних даних. Воно є багатоетапним і залежить від об’єму та якості даних. Вибір методів моделювання грає ключову роль в точності отриманих результатів.

1.1 Огляд існуючих методів

1.1.1 Методи визначення закону розподілу випадкової величини

Закон розподілу встановлює зв'язок між всіма можливими значеннями випадкової величини і відповідними ймовірностями. Знання такого закону дає можливість змоделювати та передбачити найбільш вірогідну майбутню подію.

Існує три основних сімейства методів визначення закону розподілу випадкової величин:

- Графічна оцінка;
- Параметрична оцінка;
- Непараметрична оцінка.

Графічна оцінка

Графічна оцінка полягає у візуальному аналізі графіку емпіричного закону розподілу або частіше гістограми спостережень для винесення гіпотези про густину та закон розподілу, якому може слідувати випадкова величина. Така оцінка не може дати точних даних про розподіл випадкової величини, проте дуже часто використовується для визначення можливого типу розподілу і подальшої оцінки його параметрів.

Параметрична оцінка

Параметрична оцінка полягає у визначенні параметрів розподілу що належить до певного сімейства. Модель або статистичний експеримент складається з сімейства законів розподілу $(P_\theta)_{\theta\in\Theta}$, де Θ – частина евклідового простору. Припустимо, що ці закони мають густину розподілу по відношенню до σ-скінченної міри $\mu(dx)$, що в загальному випадку являється мірою Лебега або кількісною мірою:

$$P_\theta(dy) = p_\theta(y)\mu(dy)$$

Нехай ми маємо вибірку Y, розподілену за законом P_{θ^*}, і ми шукаємо оцінку θ^*. Існує декілька основних методів оцінки :

1) Метод моментів

Він включає в себе оцінку параметрів шляхом порівняння певних теоретичних моментів (які залежать від цих параметрів) з відповідними емпіричними. Можливість порівняння випливає з закону великих чисел, що дозволяє "прирівняти" математичне очікування з середнім емпіричним. Для знаходження параметрів необхідно вирішити систему рівнянь.

Припустимо, що вибірка $X_1, \dots, X_n$ складається з незалежних, однаково розподілених спостережень і знаходиться у сім'ї параметричних законів, з параметрами θ. Будь-яка функція даних вибірки являється функцією $F(\theta)$. Це особливо вірно для моментів сімейства закону розподілу, якщо вони існують.

Ми вибираємо s моментів $G = [m_1(\theta), m_2(\theta), \dots, m_s(\theta)]$, які визначають вектор розмірністю $s \times 1$. Отже існує функція G, така що $G(\theta) = [m_1(\theta), m_2(\theta), \dots, m_s(\theta)]$. Емпіричним еквівалентом вектору G є вектор, що складається з s моментів вибірки, позначений $\widehat{G}$. Це означає, що ми замінюємо і-ий теоретичний момент $E_\theta(X^i)$ значенням:

$$\widehat{m}_i = \frac{1}{n}\sum_{k=1}^{n} x_k^i$$

Оцінка θ методом моментів позначається як $\hat{\theta}$ і знаходиться вирішенням векторного рівняння:

$$\hat{G} = G(\hat{\theta})$$

У деяких випадках, метод моментів не дістає границь Крамера-Рао: оцінка менш точна ніж при методі максимальної правдоподібності.

Тим не менш, в деяких випадках (як при гамма-розподілі), розрахунок функції правдоподібності може викликати проблеми (необхідно використовувати комп'ютер та чисельні алгоритми), у той час як метод моментів легкий у застосуванні та дає гарні результати.

Метод моментів може бути використаний в якості початкової точки для максимізації функції правдоподібності: при використанні чисельних алгоритмів, таких як метод Ньютона, необхідно задавати початкову точку.

До найзначніших мінусів відноситься те, що при не достатньо великому розмірі вибірки, закон великих чисел не діє, і, отже, емпіричні моменти не відповідають досить добре теоретичним. В такому разі, оцінка методом моментів не є достовірною: результуючі оцінювання можуть виходити за область визначення параметрів. Наприклад, для гамма-розподілу, маленький вибірка може призвести до $\alpha < 0$ [2].

2) Метод найменших квадратів

Метод найменших квадратів – метод знаходження наближеного розв'язку надлишково-визначеної системи. Зокрема важливим застосуванням є оцінка параметрів у лінійній регресії.

Для заданого набору даних $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$ будується модель:

$$y_i = \beta_0\beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \dots, n,$$

або в матричній формі:

$$y = X\beta + \varepsilon$$

де:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

В цих формулах β – вектор параметрів, які оцінюються, наприклад, за допомогою методу найменших квадратів, а ε – вектор випадкових змінних.

У класичній моделі множинної лінійної регресії приймаються такі умови:

- $y_i = \beta_0\beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i,\ i = 1, \dots, n,$
- $E[\varepsilon_i] = 0.$
- $E[\varepsilon_i\varepsilon_j] = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}$, тобто ε і є гомоскедастичними і між ними відсутня будь-яка залежність.
- Ранг матриці X рівний $p + 1$, тобто між пояснюючими змінними відсутня лінійна залежність.

Для такої моделі оцінка $\hat{\beta}$ одержана методом найменших квадратів володіє властивостями:

- Незміщеність. Оцінка $\hat{\beta}$ є незміщеною, тобто $E[\hat{\beta}|X] = \beta$. $E[\hat{\beta}] = E[(X'X)^{-1}X'(X\beta + \varepsilon)] = \beta + [(X'X)^{-1}X']E[\varepsilon] = \beta$
- Коваріаційна матриця оцінки $\hat{\beta}$ рівна: $\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

– Ефективність. Згідно з теоремою Гауса-Маркова оцінка, що одержана МНК, є найкращою лінійною незміщеною оцінкою.
– Змістовність. При доволі слабких обмеженнях на матрицю X метод найменших квадратів є змістовним, тобто при збільшенні розміру вибірки, оцінка за імовірністю прямує до точного значення параметру. Однією з достатніх умов є наприклад прямування найменшого власного значення матриці (X^TX) до безмежності при збільшенні розміру вибірки [3].

Якщо додатково припустити нормальність змінних ε, то оцінка МНК має розподіл:

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

3) Метод максимальної правдоподібності (ММП)

Це метод оцінювання невідомого параметра шляхом максимізації функції вірогідності.

Нехай маємо вибірку $X_1, \dots, X_n$ з розподілу P_θ, де $\theta \in \Theta$ – невідомий параметр. Нехай $V(x, \theta): \Theta \to \mathbb{R}$ – функція правдоподібності або вірогідності, де $x \in \mathbb{R}$. Точкова оцінка $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n) = \arg\max_{\theta \in \Theta} V(X_1, \dots, X_n, \theta)$ називається оцінкою максимальної вірогідності параметра θ. Таким чином, оцінка максимальної вірогідності – це така оцінка, яка максимізує функцію вірогідності при фіксованій реалізації вибірки.

Часто, для спрощення розрахунків використовують функцію $\ln\big(V(x, \theta)\big) = L(x, \theta)$. Оскільки функція $x \to \ln x, x > 0$, монотонно зростає на всій області визначення, максимум будь-якої функції $f(\theta)$ є максимумом функції $\ln f(\theta)$, і навпаки. Таким чином,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(X_1, \dots, X_n, \theta).$$

Функцію L також називають логарифмічною функцією вірогідності [4].

Метод оцінки максимальної вірогідності застосовується для широкого кола статистичних моделей, зокрема:

- лінійні моделі і узагальнені лінійні моделі;
- факторний аналіз;
- моделювання структурних рівнянь;
- багато ситуацій, в рамках перевірки гіпотези і довірчого інтервалу формування;
- дискретні моделі вибору.

Також він застосовується в широких областях науки, зокрема:

- системи зв'язку;
- психометрія;
- економетрика;
- час затримки в акустичних і електромагнітних системах;
- моделювання в ядерній фізиці і фізиці елементарних частинок;
- обчислювальна філогенетика;
- моделювання каналів в транспортних мережах.

До основних переваг методу відноситься точність і простота реалізації. Проте існують ситуації в яких неможливо аналітично визначити функцію вірогідності. В таких випадках необхідно використовувати чисельні методи, що породжують власні похибки і впливають на точність результату.

Непараметрична оцінка

Непараметрична оцінка має за мету через скінченну кількість спостережень оцінити невідому функцію $f \in \Theta$, де Θ – досить широкий функціональний простір.

На відміну від параметричної оцінки, результати якої найчастіше являються асимптотичними, непараметрична виводить більш конкретні результати, які можуть добре адоптуватися до реальних проблем. Але вона

має ряд своїх обмежень і складностей, через що на сьогоднішній день поки програє параметричній у використанні.

Класичними непараметричними моделями являються:

1) Модель оцінки густини розподілу

Припустимо, що маємо вибірку $X_1, \dots, X_n$ незалежних однаково розподілених випадкових величин з невідомим розподілом P_f, де f – густина розподілу, така що $f \in \mathcal{F}$, де $\mathcal{F}$ - певний функціональний простір.

Необхідно побудувати оцінку з ядром. З теореми Гливенко-Кантеллі ми маємо (рівномірно по x):

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x) \xrightarrow{\text{м.н.}} F(x) = P(X \le x),$$

де $1(\cdot)$ – характеристична функція.

Для досить малого кроку h, можна застосувати скінчено різницеву апроксимацію:

$$f(x) = F'(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

Тоді ми маємо ядерну оцінку Розенблата, що виражається як:

$$\hat{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2} 1(-h < X_i - x \le h)$$

В більш загальній формі, ядерну оцінку можна записати як:

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

де K(·) – статистичне ядро – симетрична, але не обов'язково додатна функція з інтегралом рівним одиниці (часто використовуються стандартні статистичні функції щільності), $h > 0$ – параметр згладжування, який ще називають пропускною здатністю.

Можемо показати, що при слабких припущеннях, не існує непараметричної оцінки, що сходиться швидше, ніж ядерна оцінка. Слід зазначити, що швидкість збіжності $n^{-1/5}$ нижче, ніж типова швидкість для параметричних методів, що дорівнює n^{-1}.

Практичне застосування цього методу вимагає двох речей:

– статистичне ядро K;

– параметр згладжування h.

Якщо вибір ядра не має дуже великого впливу на оцінку, це діаметрально протилежна для параметра згладжування. Занадто низьке значення h призводить до появи штучних деталей, що видно на графі оцінки. При занадто великому значенні h, більшість суттєвих деталей зникає. Тому вибір h є центральним питанням при оцінці щільності розподілу.

Поширений спосіб отримання значення h, це припустити, що вибірка розподілена відповідно до заданого параметричного закону, наприклад нормального розподілу $N(\mu, \sigma^2)$. Тоді, можна взяти оптимально пораховане значення:

$$h = 1.06\sigma n^{-1/5}$$

На жаль, оцінка гаусівським ядром не завжди є ефективною, наприклад, коли n є маленьким.

Інший спосіб підбору пропускної здатності, це пошук оптимального значення h. Нехай $R\left(f, \hat{f}(x)\right)$ – функція ризику у просторі L^2 для f. При слабких припущеннях на f і K маємо:

$$R\left(f,\hat{f}(x)\right) \approx \frac{1}{4}\sigma_K^4 h^4 \int \left(f''(x)\right)^2 dx + \frac{\int K^2(x)dx}{nh},$$

де $\sigma_K^2 = \int x^2 K(x)dx$.

Оптимальна пропускна здатність отримується мінімізації функції ризику і дорівнює:

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}},$$

де $c_1 = \int x^2 K(x)dx, c_2 = \int K^2(x)dx, c_3 = \int \left(f''(x)\right)^2 dx$, параметр h завжди пропорційний $n^{-1/5}$ [5].

2) Непараметрична регресія

Маємо вибірку $(Y_i, X_i), i = 1, \dots, n$, незалежних однаково розподілених випадкових величин, таких що:

$$Y_i = f(X_i) + \varepsilon_i,$$

де випадкові змінні відповідають вимозі $E(\varepsilon_i) = 0$ та $f \in \mathcal{F}$ – невідома функція [5].

3) Модель гаусівського білого шуму

Ми спостерігаємо траєкторію $\{Y(t), t \in [0,1]\}$ процесу Y, визначеного стохастичним диференційним рівнянням:

$$dY(t) = f(t)dt + \varepsilon dW(t), t \in [0,1],$$

де W – стандартний вінерівський процес на $[0,1]$, f – невідома функція [5].

4) Обернена статистична задача

Маємо вибірку $(Y_i, X_i), i = 1, \dots, n$, незалежних однаково розподілених випадкових величин, таких що:

$$Y_i = f(X_i, d_i) + v_i.$$

Необхідно визначити функцію розподілу X_i, опираючись на данні Y_i [5].

1.1.2 Теорія та типи копула-моделей

В теорії ймовірностей дві випадкових події називаються залежними, якщо настання одної з них змінює ймовірність настання іншої. Аналогічно, дві випадкові величини називаються залежними, якщо значення одної із них впливає на ймовірність значень іншої. В інакшому випадку і події, і величини називаються незалежними відносно ймовірності [6].

Найпершою та однією з найбільш використовуваних мір залежностей випадкових величин являється коефіцієнт кореляції, формально введений в 1888 році Френсісом Гальтоном. Проблема полягає в тому, що цей коефіцієнт є нормованою лінійною мірою залежності випадкових величин, а отже не допоможе у моделюванні складних та нелінійних типів залежностей. Тому необхідно шукати інший підхід [7].

Як було сказано раніше, для моделювання складних типів залежностей доцільно використовувати копули. Копула – це багатовимірна функція розподілу, що визначена на n-вимірному одиничному кубі $[0,1]^n$, така що кожен її маргінальний розподіл рівномірний на інтервалі [0,1].

Побудова копул базується на теоремі Склара. Вона полягає у тому, що для будь-якої двовимірної функції розподілу $H(x, y)$ з одновимірними

маргінальними функціями розподілу $F(x) = H(x, \infty)$ та $G(y) = H(\infty, y)$ існує копула, така що

$$H(x, y) = C\big(F(x), G(y)\big)$$

(де ми ототожнюємо C з її функцією розподілу)

Копула несе в собі всю інформацію про природу залежності між двома випадковими величинами, якої немає в маргінальних розподілах, але вона не містить інформації про ці розподіли. В результаті інформація про маргінали і про залежності між ними відділяються одна від одної [8].

Для визначення копули розглянемо дві випадкові величини X_1 і X_2. Обов'язкові властивості копули можуть бути сформульовані наступним чином.

1) Якщо одна із змінних копули дорівнює нулю, то значення функції дорівнює нулю:

$$\forall X_1, X_2 \in I, C(X_1, 0) = C(0, X_2) = 0$$

2) Властивість квазімонотонності. Функцію $C(X_1, X_2)$ будемо називати квазімонотонною, якщо виконується наступне:

$$\forall X_1(1) \le X_1(2), X_2(1) \le X_2(2), B = [X_1(1), X_1(2)] \times [X_2(1), X_2(2)]$$

$$V_c(B) = C\big(X_1(2), X_2(2)\big) - C\big(X_1(1), X_2(2)\big) - C\big(X_1(2), X_2(1)\big) + C\big(X_1(1), X_2(1)\big) \ge 0$$

3) Якщо одна із змінних копули дорівнює одиниці, то значення копули дорівнює значенню другої змінної:

$$\forall X_1, X_2 \in I, C(X_1, 1) = X_1, C(1, X_2) = X_2$$

Представляється можливим знайти найменшу за значенням і найбільшу з усіх можливих копули. Тому для копули завжди виконується умова дотримання границь Фреше-Хефдінга:

$$\max(0, X_1 + X_2 - 1) \leq C(X_1, X_2) \leq \min(X_1, X_2)$$

Межі Фреше-Хефдінга обмежують будь-яку копулу знизу мінімальної копулою, а зверху максимальною копулою. Мінімальна копула (рис. 1.1), виражена функцією $C(X_1, X_2) = \max(0, X_1 + X_2 - 1)$, є найменшою функцією, що задовольняє всім умовам копулярної функції. Максимальна копули (рис. 1.2) має вигляд $C(X_1, X_2) = \min(X_1, X_2)$ і є найбільшою такою функцією [9].

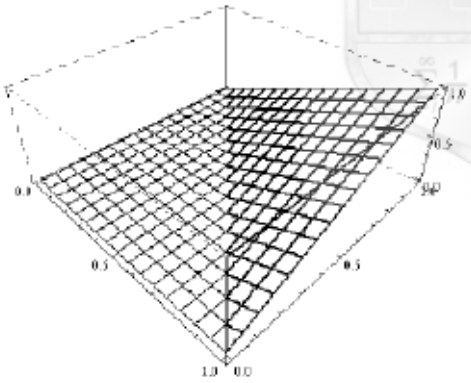

Рисунок 1.1. Мінімальна копула

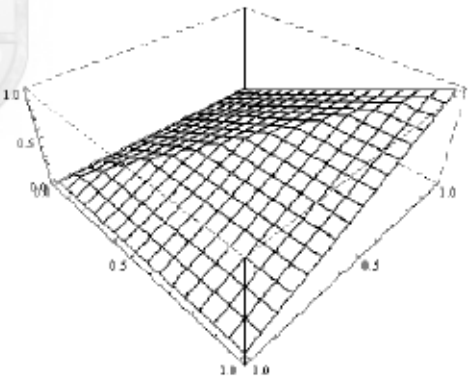

Рисунок 1.2. Максимальна копула

Копула може бути незалежною і повноцінною. Незалежні копули, або копули добутку, називають копули, які характеризують випадкові величини як незалежні одна від одної і визначаються як добуток аргументів функції:

$$C(X_1, X_2) = X_1 \cdot X_2$$

Повноцінною копулою є копула, яка включає границі Фреше-Хефдінга, а також випадок незалежної копули, як окремі випадки, визначені деякими значеннями параметрів.

Копулярні функції поділяються на сімейства. Основними сімействами копул є: еліптичні, архімедові та копули екстремальних значень. Також, як окремі сімейства, можна розглянути емпіричні копули та копули, що не відносяться до жодного іншого сімейства [10].

Еліптичні копули

Основою для еліптичної копули служать аналітичні записи багатовимірного розподілу Гауса і Стьюдента відповідно, що дозволяє створити симетричні спільні розподіли. Для отримання багатовимірного нормального розподілу використовуються гаусівська копула і гаусівські маргінальні розподіли. Аналогічно, для отримання розподілу Стьюдента використовується копула Стьюдента та маргінальні розподіли Стьюдента з однаковим числом ступенів свободи.

Копула Гауса:

$$C\left(u^{(1)}, u^{(2)}\right) = \int_{-\infty}^{u^{(2)}} \int_{-\infty}^{u^{(1)}} \frac{1}{2\pi\sqrt{1-\alpha^2}} \exp\left(\frac{2\alpha z_1 z_2 - z_1^2 z_2^2}{2(1-\alpha^2)}\right) dz_1\, dz_2$$

де α – параметр копули, а $u^{(1)}, u^{(2)}$ розподілені за нормальним розподілом Гауса:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(\frac{(t-m)^2}{2\sigma^2}\right) dt$$

Гаусівські копули мають достатньо обмежене застосування. Так, у одній з робіт Олександра Карола та Жака Пезьє стверджується, що

використання копули Гауса допустимо лише у випадках, коли залежність випадкових величин може бути достатньою мірою змодельована лінійної кореляцією [11].

Копула Стьюдента:

$$C\left(u^{(1)}, u^{(2)}\right) = \int_{-\infty}^{u^{(2)}} \int_{-\infty}^{u^{(1)}} \frac{1}{2\pi\sqrt{1-\alpha^2}} \exp\left(1 + \frac{z_1^2 + z_2^2 - 2\alpha z_1 z_2}{d(1-\alpha^2)}\right)^{\frac{d+2}{2}} dz_1\, dz_2$$

де α – параметр копули, d – число ступенів свободи, а $u^{(1)}, u^{(2)}$ розподілені за розподілом Стьюдента:

$$F_d(x) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi d}\,\Gamma\left(\frac{d}{2}\right)} \int_{-\infty}^{x} \exp\left(1 + \frac{t^2}{d}\right)^{-\frac{d+1}{2}} dt$$

Архімедові копули

Одними з найбільш поширених копул є архімедові копули, які задаються формулою:

$$H(x,y) = \Psi^{-1}\left(\Psi\left(F(x)\right) + \Psi\left(G(y)\right)\right)$$

де Ψ – функція генератор, а F, G – функції розподілу випадкових величин X та Y відповідно.

Для правильної побудови копули, функція Ψ має відповідати наступним умовам:

а) $\Psi(1) = 0$;

б) $\lim\limits_{x\to\infty} \Psi(x) = \infty$;

в) $\Psi'(x) < 0$;

г) $\Psi''(x) > 0$.

Найпопулярнішими з архімедових копул є:

1) Копула Клейтона

$$\Psi(x) = x^{\theta} - 1;\ \ \theta \le 0; H(x,y) = \left(F(x)^{\theta} + G(y)^{\theta} - 1\right)^{\frac{1}{\theta}}$$

2) Копула Гамбела

$$\Psi(x) = (-\ln(x))^{\alpha};\ \ H(x,y) = \exp\left(-\left(\left(-\ln\left(F(x)\right)\right)^{\alpha} + \left(-\ln\left(G(y)\right)\right)^{\alpha}\right)^{\frac{1}{\alpha}}\right)$$

3) Копула Франка

$$\Psi_{\alpha}(x) = -\ln\left(\frac{\exp(-x\alpha) - 1}{\exp(-\alpha) - 1}\right);$$

$$H(x,y) = -\frac{1}{\alpha}\ln\left(1 + \frac{\left(e^{-\alpha F(x)} - 1\right)\left(e^{-\alpha G(y)} - 1\right)}{\left(e^{-\alpha} - 1\right)}\right)$$

Копули екстремальних значень

Нехай $(X_{i1}, \dots, X_{im})$, $i = 1,2, \dots$ є незалежними однаково розподіленими m-вимірними випадковими векторами з функцією розподілу F.

Нехай

$$M_{ij} = \max_{1 \le i \le n} X_{ij}, j = 1, \dots, m$$

є по компонентним максимумом. Багатовимірний розподіл екстремальних значень є границею випадкових векторів

$\left(\frac{(M_{1n}-a_{1n})}{b_{1n}}, \dots, \frac{(M_{mn}-a_{mn})}{b_{mn}}\right)$. Якщо граничний розподіл існує, то кожна одновимірна компонента розподілу є одновимірним розподілом екстремальних значень, який можна записати у вигляді:

$$C\big(H(z_1;\gamma_1), \dots, H(z_m;\gamma_m)\big)$$

де $H\big(z_j;\gamma_j\big)$ є узагальненим розподілом екстремального значення, а C є копулою.

Для копул такого типу необхідне виконання умови:

$$C(X_1^t, X_2^t) = C^t(X_1, X_2)$$

Характерними представниками копул екстремальних значень являються:

1) Копула Галамбоса

$$C(X_1, X_2) = X_1 X_2 \exp\left(((-\ln(X_1))^{-\alpha} + (-\ln(X_2))^{-\alpha})^{\frac{1}{\alpha}}\right)$$

2) Копула Хаслера-Рейса

$$C(X_1, X_2) = \exp\left(-X_1\left(\frac{1}{\alpha} + \frac{\alpha}{2}\ln\frac{\ln X_1}{\ln X_2}\right) - X_2\left(\frac{1}{\alpha} + \frac{\alpha}{2}\ln\frac{\ln X_2}{\ln X_1}\right)\right)$$

Емпірична копула

При аналізі даних з невідомим розподілом, можна побудувати "емпіричну копулу" шляхом такої згортки, щоб маргінальні розподіли вийшли рівномірними. Математично це можна записати так:

$$C_n\left(\frac{i}{n},\frac{j}{n}\right)=\frac{1}{n}$$

Число пар (x,y) таких що $x \le x_{(i)}, y \le y_{(i)}, 1 \le i \le n, 1 \le j \le n$, де $x_{(i)}$ – і-та порядкова статистика x.

Інші копули

В окрему категорію можна винести такі функції, які не відносяться ні до одного з вище вказаних сімейств. Так, копули Плаке і Фарлі-Гамбела-Моргенштерна (FGM) не є копулами ні еліптичного класу, ні копулами екстремальних значень. В їх функціональній формі неможливо виділити функцію-генератор, як в архімедових. Копула Плаке була утворена алгебраїчним способом, коли на основі аналітичних таблиць обчислювалася ступінь залежності стовпців від строк. Копула Фарлі-Гумбеля-Моргенштерна, маючи досить просту функціональну форму, дозволяють моделювати несильну ступінь залежності. Функція копули FGM має наступний вигляд:

$$C_\alpha^{FGM}(X_1,X_2)=X_1X_2\big(1+\alpha(1-X_1)(1-X_2)\big)$$

Методи оцінки параметрів копула-моделей

1) Параметричні (MLE, IFM)

Цей клас методів припускає параметризацію як граничних розподілів, так і зв'язки. Якщо базовий підхід метод найбільшої правдоподібності (англ. Maximum Likelihood Estimation) передбачає максимізацію функції правдоподібності одночасно по граничних розподілах і по зв'язці, то метод «від маргіналів» (Inference for Margin — IFM) передбачає два етапи оцінки: спочатку — параметризація граничних розподілів, потім — копули.

2) Напівпараметричні (SP, CML)

Напівпараметричні методи також припускають двоетапну оцінку копули. Але на першому етапі замість оцінки граничних розподілів

використовується емпіричний розподіл. На другому ж етапі відбувається параметрична оцінка копули. Було показано, що напівпараметричний метод (SP — semi-parametric) дає більш ефективні і стійкі оцінки ніж параметричні методи у випадках, коли тип оцінюваного розподілу не відомий і, як наслідок, виникає загроза їхньої невірної специфікації [12].

3) Непараметричні

Серед непараметричних методів оцінки копул можна виділити підходи на основі оцінки емпіричної копули і ядерних оцінок. Перший підхід передбачає оцінку функції розподілу емпіричної копули, що відображає кількість випадків, коли реалізації випадкових величин одночасно потрапили в обрану групу розбиття нескінченного ймовірнісного простору.

1.1.3 Моделювання хвостів розподілу

Моделювання хвостів розподілів полягає у визначенні функції розподілу екстремальних значень, тобто таких значень, що перевищують заданий поріг. Розподілами екстремальних значень спочатку цікавилися фахівці з «абстрактної» теорії ймовірностей, та фахівці в прикладних областях - інженери та гідрологи. Тільки з недавніх пір ці розподілу увійшли в сферу істотних інтересів фахівців по статистиці.

Вивчення властивостей розподілів екстремальних значень протягом довгого часу перебувало трохи осторонь від основних напрямків статистичної теорії розподілів. Справа в тому, що на ранній стадії створення статистичної теорії основна увага приділялася проблемам підгонки кривих розподілу, і лише значно пізніше - розвитку теорії статистичного висновку. В даний час теорія розподілу екстремальних значень є складовою частиною багатьох природничо-наукових дисциплін. Основними областями застосування є: промисловість (визначення надійності конструкцій),

страхування (особливо перестрахування), фінансовий аналіз (наприклад, для прогнозування краху фондового ринку або валютних криз), вивчення таких явищ як зливи, урагани, повені, забруднення атмосфери і корозія, а також тонкі математичні результати, що стосуються точкових випадкових процесів і регулярно мінливих функцій.

Практична цінність такого моделювання полягає в тому, щоб оцінити вартість ризику для подій, що мають маленьку ймовірність настання, проте якою не можна нехтувати, тобто, рідкісних подій, що мають велике значення.

Перша теорема теорії екстремальних значень

Теорія екстремальних значень була розроблена для оцінки ймовірності виникнення рідкісних подій. Вона дозволяє екстраполювати поведінку хвоста розподілу даних через найбільші з спостережуваних даних (екстремальні дані).

В основі теорії лежить теорема про розподіл екстремальних значень (або EVD, Extreme Value Distribution), що для максимуму з n спостережень, є аналогом центральної граничної теореми для середнього значення. Вона описує можливі рамки закону максимуму з n спостережень, правильно нормованого за допомогою двох послідовностей $(\alpha_n)_{n\geq 1}$ і $(\beta_n)_{n\geq 1}$, з n незалежних і однаково розподілених випадкових величин. Якщо F - функція розподілу загального закону цих n випадкових величин, закон розподілу їх максимума - F^n .

Іншими словами, якщо F – функція розподілу закону, що нас цікавить, за певних умов регулярності F, існують $\tau \in \mathbb{R}$ і дві дійсні нормалізуючі послідовності $(\alpha_n)_{n\geq 1}$ та $(\beta_n)_{n\geq 1}$ $(\beta_n > 0)$, такі що

$$\forall x \in \mathbb{R}, \qquad \lim_{n\to\infty} F^n(\alpha_n + \beta_n x) = H_\tau(x)$$

де H_τ – функція розподілу закону екстремальних значень:

$$H_\tau(x) := \begin{cases} \exp\left[-(1+\tau x)^{-\frac{1}{\tau}}\right], & \forall\, x, 1+\tau x > 0, \quad \text{якщо } \tau \neq 0 \\ \exp[-e^{-x}], & \forall x \in \mathbb{R}, \quad \text{якщо } \tau = 0 \end{cases}$$

Коли $\tau \neq 0$ і $1 + \tau x \leq 0, H_\tau(x) = 0$ [13].

Також цю теорему називають Першою теоремою теорії екстремальних значень.

Кажуть, що функція розподілу F (або відповідний закон) знаходиться в області притягання (Domain of Attraction, DA) Гамбела (тип 1), Фреше (тип 2), і Вейбулла (тип 3) залежно від того, що $\tau = 0$, $\tau > 0$, або $\tau < 0$. Позначимо ці області відповідно DA(Гамбела), DA(Фреше) і DA(Вейбулла). Кінцевою точкою функції розподілу F називають дійсне $\omega(F)$ (скінченне або нескінченне), що визначається як $\omega(F) := \sup[x: F(x) < 1]$.

Можна підсумувати, що:

- для законів з DA(Фреше), таких як закон Стьюдента, кінцева точка нескінченна $\omega(F) = +\infty$. Це свідчить про те, що розподіл максимумів F^n має важкий хвіст (включаючи поліноміальний розпад);
- для законів з DA(Вейбулла), такі як закон Бета розподілу, кінцева точка завжди звичайно $\omega(F) < +\infty$. Це свідчить про те, що розподіл максимумів F^n має легкий хвіст з скінченною верхньою межею;
- для законів з DA(Гамбела), кінцева точка може бути скінченною або нескінченною, наприклад, для нормального закону, логнормального, Гамма і Вейбулла. Це свідчить про те, що розподіл F^n має експоненційний хвіст [14].

Друга теорема теорії екстремальних значень

Другий способів оцінки хвостів розподілу є метод ексцесів (також відомий як POT, Peaks Over Threshold). Візьмемо дійсне u "досить велике" але менше ніж максимальне значення розподілу ($u < \omega(F)$) і назвемо його

порогом. Метод ексцесів базується на апроксимації закону, за яким розподіленні значення випадкової величини X, що перевищують поріг u, тобто, умовному розподілі дійсної позитивної випадкової величини $X - u$, знаючи, що $X > u$. Функція розподілу значень, що перевищують поріг u визначається як:

$$F_u(y) = P(X - u \le y | X > u) = \begin{cases} \dfrac{F(u+y) - F(u)}{1 - F(u)}, & y \ge 0 \\ 0, & y < 0 \end{cases}$$

Асимптотичний закон ексцесів (або надлишків), показаний теоремою, що довів Джеймс Пікандс в 1975 році. Ця теорема дає асимптотичний розподіл хвостів випадкової величини X, коли дійсний розподіл F невідомий.

Нехай ми маємо $(X_1, X_2, \ldots)$ послідовність незалежних однаково розподілених випадкових величин, що розподілені за законом, який належить до одного з типів областей притягання, і F_u – функція їх умовного розподілу надлишкових величин, то при великих значеннях u, F_u може бути гарно апроксимоване узагальненим розподілом Парето. Тобто:

$$F_u(y) \to G_{k,\sigma}(y), \qquad \text{коли } u \to \infty$$

де $G_{u,\sigma}(y)$ – Узагальнений розподіл Парето з функцією розподілу:

$$G_{u,\mu,\sigma}(y) = \begin{cases} 1 - \left(1 + \dfrac{u(y-\mu)}{\sigma}\right)^{-\frac{1}{u}}, & \text{якщо } u \ne 0 \\ 1 - e^{-\frac{(y-\mu)}{\sigma}}, & \text{якщо } u = 0 \end{cases} \qquad (1)$$

Тут $\sigma > 0$ і $y \ge 0$ коли $u \ge 0$ та $0 \le y \le -\frac{\sigma}{u}$ коли $u < 0$ [15].

1.1.4 Методи перевірки гіпотез та якості моделей

Будь-які результати та гіпотези отримані при статистичному аналізі, мають бути перевірені, на скільки вони відповідають реальності. Для цього існує багато різноманітних тестів і критеріїв, що мають свої властивості та особливості.

Так як основною ціллю даної роботи являється побудова копула-моделі, необхідно буде оцінити якість отриманого результату. Найбільш розповсюдженим критерієм вибору оптимальної копули є критерій на основі значення функції максимальної правдоподібності – критерії Акайке (AI), Шварца (BI) та співставлення коефіцієнта Кендалла з модельним значенням τ (тау). Другими за частототою застосування є тести Колмогорова-Смирнова й Андерсона-Дарлінга. Третім є метод оцінки дистанції до емпіричної копули.

Графічна перевірка відповідності обраного розподілу

Це найбільш інтуїтивний та простий метод для перевірки чи гарно був підібраний розподіл та оцінені параметри. Він полягає в співставленні та порівнянні емпіричної функції розподілу або гістограми значень з певною теоретичною функцією розподілу або густиною ймовірності відповідно. Така перевірка не може бути дуже точною, проте вона може одразу показати недоліки обраної моделі.

Непараметричний тест Колмогорова-Смирнова

У статистиці, тест Колмогорова-Смирнова є критерієм перевірки гіпотези, що використовується для визначення, чи розподілена випадкова величина за відомим законом, даним його безперервною функцією розподілу, або для перевірки, чи розподілені дві вибірки за однаковим законом.

Цей тест базується на властивостях емпіричної функції розподілу: якщо ми маємо вибірку $X_1, \dots, X_n$ з n дійсних незалежних випадкових величин, то емпірична функція розподілу цієї вибірки визначається як:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x),$$

де $1(X_i \le x) = \begin{cases} 1, & \text{якщо } X_i \le x \\ 0, & \text{інакше} \end{cases}$

Статистика критерію для емпіричної функції розподілу $\hat{F}_n(x)$ визначається наступним чином:

$$D_n = \sup_x \left|\hat{F}_n(x) - F(x)\right|,$$

де $\sup S$ – точна верхня межа множини $S = \left|\hat{F}_n(x) - F(x)\right|$, а F – модель.

Позначимо нульову гіпотезу H_0, як гіпотезу про те, що вибірка слідує розподілу $F(x)$, тоді по теоремі Колмогорова для введеної статистики справедливо:

$$\forall t > 0: \lim_{n\to\infty} P\left(\sqrt{n}D_n \le t\right) = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}$$

Це робить його одним з основних кандидатів для тест гіпотез, щоб визначити, чи дійсно слідує вибірка обраному закону, або дві вибірки мають однаковий закон, якщо їх функції розподілу безперервні [16].

Також даний тест часто застосовують для перевірки якості генераторів випадкових чисел.

У сучасних програмних реалізаціях цього тесту результатами являються дві величини: статистика $D = \sup_x \left|\hat{F}_n(x) - F(x)\right|$ та P-value. Ці результати можна інтерпретувати таким чином: в разі прийняття нульової гіпотези H_0 (дів вибірки розподілені за одним законом, або вибірка слідує обраному закону), максимальна різниця між функціями розподілу буде

дорівнювати D з ймовірністю P-value. Тобто, чим менше D та більше P-value, тим більша ймовірність прийняту H_0.

Тест Андерсона-Дарлінга

Класичний непараметричний критерій згоди Андерсона-Дарлінга [1, 2] призначений для перевірки простих гіпотез про приналежність аналізованої вибірки повністю відомому закону (про згоду емпіричного розподілу $\hat{F}_n(x)$ і теоретичного закону $F(x, \theta)$, тобто для перевірки гіпотез вигляду $H_0: \hat{F}_n(x) = F(x, \theta)$ з відомим вектором параметрів теоретичного закону.

У критерії Ω^2 Андерсона-Дарлінга використовується статистика:

$$S_\Omega = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln\left(F(x_i,\theta)\right) + \left(1 - \frac{2i-1}{2n}\right)\ln\left(1 - F(x_i,\theta)\right)\right\},$$

де n – обсяг вибірки, $x_1, \dots, x_n$ – впорядковані за зростанням елементи вибірки.

При справедливості простої нульової гіпотези, статистика критерію підкоряється розподілу виду $\alpha^2(S)$.

При перевірці простих гіпотез критерій є вільним від розподілу, тобто не залежить від виду закону, з яким перевіряється узгодженість.

При перевірці складних гіпотез виду $H_0: \hat{F}_n(x) \in \{F(x, \theta), \theta \in \Theta\}$, де оцінка $\hat{\theta}$ скалярного або векторного параметра розподілу $F(x, \theta)$ обчислюється за тією ж самою вибіркою, непараметричні критерії згоди втрачають властивість свободи від розподілу (розподілом статистики при справедливості H_0 вже не буде розподіл $\alpha^2(S)$.

При перевірці складних гіпотез розподілу статистики непараметричних критеріїв згоди залежать від ряду чинників: від виду спостережуваного закону $F(x, \theta)$, відповідного справедливій гіпотезі H_0; від типу оцінюваного параметра і числа оцінюваних параметрів; в деяких випадках від конкретного значення параметра (наприклад, у випадку сімейств гамма- і бета-розподілів);

від методу оцінювання параметрів. Відмінності в граничних розподілах тієї ж самої статистики при перевірці простих і складних гіпотез настільки істотні, що нехтувати цим ні в якому разі не можна [17].

Критерії Акаіке та Шварца

Інформаційний критерій Акаіке (AIC) - критерій, що застосовується виключно для вибору з декількох статистичних моделей.

У загальному випадку AIC:

$$AIC = 2k - 2\ln(L)$$

де k - число параметрів у статистичній моделі, і L - максимізоване значення функції правдоподібності моделі.

Уявімо, що помилки моделі нормально і незалежно розподілені. Нехай n – число спостережень і RSS – залишкова сума квадратів, визначена як:

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Припустимо, що дисперсія помилок моделі невідома, але однакова для всіх них. Отже:

$$AIC = 2k + n\left[\ln\left(2\pi\frac{RSS}{n}\right) + 1\right]$$

У разі порівняння моделей на вибірках однакової довжини, вираз можна спростити, викидаючи члени, що залежать тільки від n:

$$AIC = 2k + n[\ln(RSS)]$$

Таким чином, критерій не тільки винагороджує за якість наближення, але і штрафує за використання зайвої кількості параметрів моделі. Вважається, що найкращою буде модель з найменшим значенням критерію AIC. Критерій Шварца (BIC) штрафує вільні параметри в більшій мірі.

Варто відзначити, що абсолютне значення AIC не має сенсу - він вказує лише на відносний порядок порівнюваних моделей [18].

Тест на основі коефіцієнта Кендала

Коефіцієнт конкордації Кендалла для двох рядів X і Y довжиною n розраховується наступним чином:

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i} \operatorname{sign}\left[(x_i - x_j)(y_i - y_j)\right]$$

де $x_i, x_j \in X$ і $y_i, y_j \in Y$.

Довірчий інтервал для коефіцієнта Кендалла розраховується за формулою:

$$\hat{\tau} - u_{\frac{1+p}{2}} \sqrt{\frac{2}{n}(1-\hat{\tau}^2)} < \tau < \hat{\tau} + u_{\frac{1+p}{2}} \sqrt{\frac{2}{n}(1-\hat{\tau}^2)}$$

де u_q – q-квантиль стандартного нормального розподілу.

Розраховані значення коефіцієнта Кендалла порівнюються з модельним значенням τ, яке розраховується на основі оціненого параметра копули α. Для кожної копули визначена своя формула розрахунку модельного тау [19].

1.2 Огляд та первинний аналіз вхідних даних

Для побудови моделі було використано реальні історичні дані страхової компанії «Х», по застрахованим від пожежі будівлям. Вони представляють собою таблицю з 2167 спостережень та з наступними колонками:

1. Дата (date);
2. Збитки нанесені будівлі (building);
3. Збитки нанесені майну та матеріалам (contents);
4. Втрачений прибуток від використання будівлі (profits);
5. Сумарні збитки спричинені пожежею (total).

Збитки в таблиці наведені в млн. данських крон (1 данська крона дорівнює приблизно 0,13 євро).

Щоб отримати уявлення про випадкові величини, які будуть змодельовані в даній роботі необхідно провести аналіз основної статистичної інформації вхідних даних.

Таблиця 1.1 – Основна статистична інформація початкових ВВ

Випадкова величина	Кількість спостереж.	Середнє значення	Медіана	Середнє квадрат. відхилення	MIN	MAX	Квантиль 25%	Квантиль 75%
building	2167	1.8244	1.2701	4.3607	0	152.4132	0.854	1.9
contents	2167	1.3185	0.3727	4.7601	0	132.0132	0.0723	1.1123
profits	2167	0.2421	0	1.6167	0	61.9327	0	0.059
total	2167	3.385	1.7782	8.5075	1	263.2504	1.3211	2.9670

З табл. 1.1 одразу можна помітити, що у всіх випадках максимальне значення випадкових величин значно більше за медіану та середньоквадратичне відхилення більше за середнє значення. Це є явною ознакою наявності важких хвостів у розподілах цих випадкових величин.

Так як нас цікавить побудова двовимірної копули, доцільно взяти першу випадкову величину buil = building, а другу other = contents + profits. Очевидно, що вони є залежними і можна справедливо припустити, що вони мають схожі розподіли.

Враховуючи, що мінімальне значення суми збитків дорівнює 1 млн. данських крон та курс цієї валюти, було вирішено відібрати для розрахунків лише страхові випадки ціною більше 1 млн. по кожному параметру (327 спостережень). Основні статистичні дані нових випадкових величин наведені в табл. 1.2

Таблиця 1.2 – Основна статистична інформація синтизованих ВВ

Випадкова величина	Кількість спостереж.	Середнє значення	Медіана	Середнє квадрат. відхилення	MIN	MAX	Квантиль 25%	Квантиль 75%
buil	327	3.0712	1.5157	6.5388	0.0107	94.1684	0.6308	3.2337
other	327	5.2551	1.7829	13.5502	0.0011	167.082	0.5739	4.8865

Також важливо відмітити, що ці пожежі відбулися за проміжок часу від 03/01/1980 до 31/12/1990. Тобто дані зібрані за 10 років.

1.3 Висновки до першого розділу

В цьому розділі були розглянуті статистичні методи, що необхідні для побудови копула моделі та дані на основі яких буде базуватися моделювання.

Основні методи для визначення функції розподілу випадкової величини на основі даних вибірки поділяються на параметричні та непараметричні. В цій роботі нас більше будуть цікавити параметричні, так як вони легші у реалізації та дають достатньо точні результати.

Для побудови копули на сонові маргінальних розподілів, будуть взяті основні представники всіх трьох сімейств копул: еліптичних, архімедових та екстремальних. Результати будуть порівняні на основі коефіцієнту конкордації Кендала та буде обрана найкраща модель.

Для моделювання важких хвостів розподілу, ми скористаємося теоремою Пікандса та узагальненим розподілом Парето.

2 ПОБУДОВА КОПУЛ

2.1 Визначення законів розподілу buil та other

Першим етапом побудови копула-моделі являється визначення її маргінальних розподілів. Як було показано вище, випадкові величини buil та other мають деякі статистичні дані, що можуть підказати нам якому розподілу вони слідують. По-перше, у двох випадкових величинах, середньоквадратичне відхилення значно більше за математичне очікування і максимуми знаходяться далеко від середнього значення. Це означає, що присутня сильна асиметрія в сторону великих значень, та ймовірна наявність важких хвостів справа. С цього випливає, що можливі кандидати на роль маргінальних розподілів закони Парето та логнормальний. По-друге, медіана знаходиться не далеко від середнього значення. Це може вказувати на Гама-розподіл. Також в якості наглядного прикладу, візьмемо розподіл Вейбула, що має відмінні характеристики, а отже покаже гірші результати [20].

Розподіл Парето:

$$F_X(x) = P(X < x) = 1 - \left(\frac{x_m}{x}\right)^k, \forall x \geq x_m, \text{ де } x_m, k > 0.$$

Густина розподілу Парето має вигляд:

$$f_X(x) = \begin{cases} \frac{k x_m^k}{x^{k+1}}, x \geq x_m \\ 0, \ x < x_m \end{cases}$$

Логнормальний розподіл:

Нехай розподіл випадкової величини X задається густиною ймовірності, що має вигляд:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}},$$

де $x > 0, \sigma > 0, \mu \in \mathbb{R}$. Тоді кажуть, що X має логнормальний розподіл з параметрами σ і μ.

В дійсності, $X \sim \ln N(\mu, \sigma^2)$.

Розподіл Вейбула:

Функція розподілу випадковою величини X задається формулою:

$$F_X(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}$$

Густина розподілу має вигляд:

$$f_X(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, x \geq 0 \\ 0, \quad x < 0 \end{cases}$$

Гама-розподіл:

Нехай розподіл випадкової величини X задається густиною ймовірності, що має вигляд:

$$f_X(x) = \begin{cases} \frac{x^{k-1} e^{-\left(\frac{x}{\theta}\right)}}{\theta^k \Gamma(k)}, x \geq 0 \\ 0, \quad x < 0 \end{cases}$$, де $\Gamma(k)$ – Гамма функція Ейлера

Закон розподілу має вигляд: $F_X(x) = \frac{\gamma\left(x\frac{k}{\theta}\right)}{\Gamma(k)}$

За допомогою ММВ та статистичної мови програмування R оцінимо параметри обраних розподілів. Результати наведені в табл. 2.1.

Таблиця 2.1 – Оцінені параметри

	Парето		Логнормальний		Вейбул		Гама	
	x_m	k	μ	σ	λ	k	θ	k
buil	2.6933	5.0517	0.2720	1.3907	0.7823	2.6820	0.7096	0.2311
other	1.3429	2.4800	0.4011	1.7456	0.6314	3.4311	0.5042	0.0960

Щоб отримати уявлення, який з розподілів з оціненими параметрами краще відповідає реальним даним, побудуємо суперпозиції графіків теоретичних законів розподілів до емпіричної функції розподілу та функції густини розподілів до гістограми спостережень. Результати представлені на рис. 2.1. та 2.2.

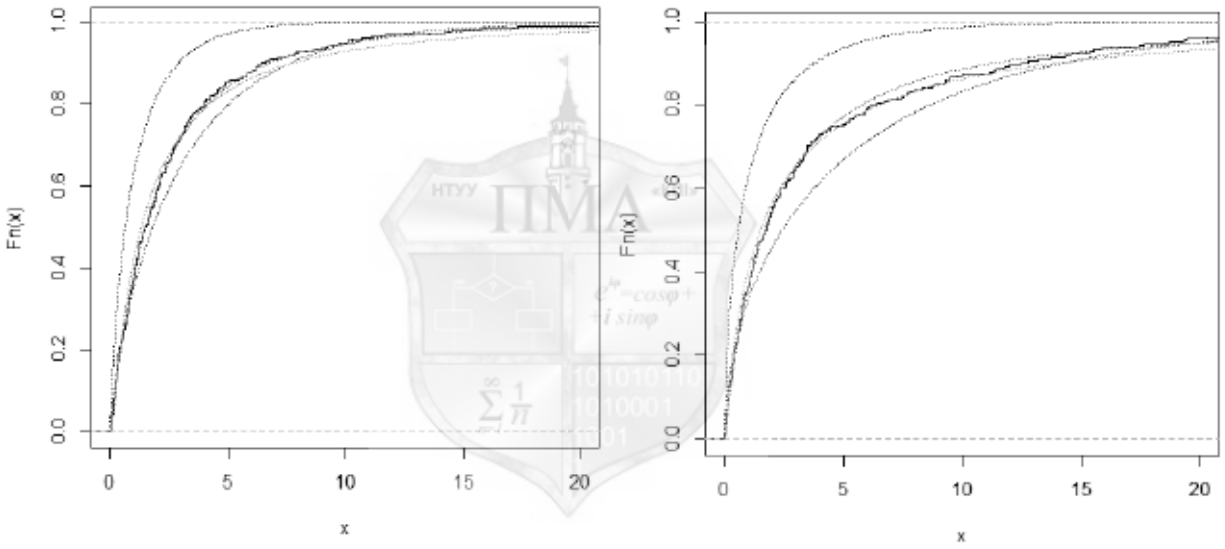

Рисунок 2.1. Суперпозиція функцій розподілів

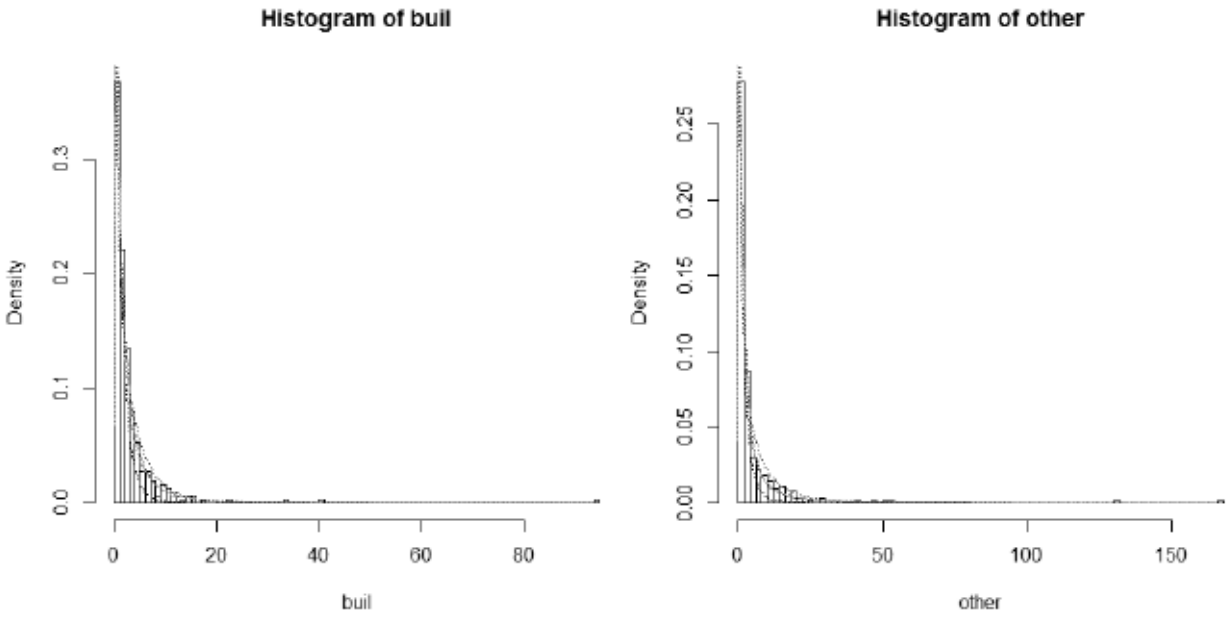

Рисунок 2.2. Суперпозиція густин розподілів

На графіках емпіричні дані зображені чорним кольором, функції розподілу та густини для Гама-розподілу – червоним, для Вейбула – синім, для Логнормального – зеленим і для Парето – фіолетовим. На рисунках добре видно, що розподіл Вейбула дуже далеко від дійсності, а Гама-розподіл, добре підходить на хвостах, але гірше на тілі розподілу. Очевидно, логнормальний розподіл та розподіл Парето найкраще підходять для описання наших спостережень, як у випадку BB buil так і у випадку other. Для підтвердження цієї гіпотези, скористаємося КС-тестом. Результати наведені в табл. 2.2.

Таблиця 2.2 – Результати КС-тесту

	Парето		Логнормальний		Вейбул		Гама	
	D	P-value	D	P-value	D	P-value	D	P-value
buil	0.0298	0.9333	0.0621	0.1605	0.067	0.1061	0.0819	0.02497
other	0.0291	0.9451	0.0572	0.2348	0.0674	0.1021	0.1198	0.0001682

Очевидним переможцем являється розподіл Парето для обох ВВ. Так як його значення статистики D найменше, з найбільшою ймовірністю P-Value.

Тобто визначені маргінальні розподіли для копула-моделі, що представляють собою:

- Випадкова величина buil розподілена за законом розподілу Парето з параметрами: $x_m = 2.6933, k = 5.0517$;
- Випадкова величина other розподілена за законом розподілу Парето з параметрами: $x_m = 1.3429, k = 2.4800$.

2.2 Вибір копули та оцінка параметрів

Кожна копула, що належить до того чи іншого сімейства, має свої особливості описання залежності між ВВ. Саме тому для побудови найкращої копула-моделі необхідно, протестувати всі сімейства копул. Це допоможе нам краще зрозуміти, яка залежність поєднує нашу модель.

Було вирішено взяти наступні копули для побудови моделі:

- Гаусівську та Стьюдента (еліптичні копули);
- Франка та Гамбела (архімедові копули);
- Хаслера-Райса (екстремальна копула).

Для початку, знайдемо коефіцієнт лінійної кореляції між ВВ buil та other. Він розраховується за формулою:

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \sum(Y - \overline{Y})^2}}$$

В нашому випадку, ми оперуємо емпіричними даними, тому формула набуває вигляду:

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \widehat{X})(Y_i - \widehat{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \widehat{X})^2 \sum_{i=1}^{n}(Y_i - \widehat{Y})^2}}$$

Результатом, отриманим за допомогою мови програмування R являється: $r_{buil,other} = 0.6303464$.

Тепер, за допомогою ММП необхідно оцінити параметри копул, з відомими маргінальними розподілами. Результати наведені в табл. 2.3.

Таблиця 2.3 – Параметри копула-моделей

Копула	Гаусівська	Стьюдента	Франка	Гамбела	Хаслера-Рейса
Параметр α	0.3632	[0.3630; 1000.0]	2.0222	1.2751	0.9372

Отже ми маємо 5 копула-моделей, що описують залежність buil і other. Необхідно вибрати найкращу з них. Для цього ми скористаємося методом перевірки якості копули на основі коефіцієнту конкордації Кендала. Розраховані значення тау, наведені в табл. 2.4.

Таблиця 2.4 – Порівняння якості копула-моделей

Копула	Емпірична	Гаусівська	Стьюдента	Франка	Гамбела	Хаслера-Рейса
τ (тау)	0.2259	0.2366	0.2366	0.2161	0.2158	0.2285

Порівнявши ці значення, ми можемо зробити висновок, що найкраща копула-модель побудована на основі копули Хаслера-Рейса (тау конкордації емпіричних даних найближче до теоретичного тау копули). Також непоганий результат показала копула Франка. Це пов'язано з тим, що в даних явно присутній важких хвіст на великих значеннях.

2.3 Висновки до другого розділу

В цьому розділі було визначено, що випадкові величини buil і other розподілені за законом Парето, та залежність що існує між ними найкращим чином пояснюється за допомогою копули екстремальних значень Хаслера-Рейса.

Були оцінені параметри, та отримана шукана модель:

- $X_{buil} \sim \text{Pareto}(x_m = 2.6933, k = 5.0517)$;
- $X_{other} \sim \text{Pareto}(x_m = 1.3429, k = 2.4800)$;
- Залежність описана копулою Хаслера-Рейса з параметром $\alpha = 0.9372$ і представлена як :
-

$$C(X_{buil}, X_{other}) = \exp\left(-X_{buil}\left(4.3764 + 0.1143 \ln \frac{\ln X_{buil}}{\ln X_{other}}\right) - X_{other}\left(4.3764 + 0.1143 \ln \frac{\ln X_{other}}{\ln X_{buil}}\right)\right)$$

Це підтверджує той факт, що у ВВ дуже важливу роль грають екстремальні значення. Тому необхідно перевірити, як гарно ця модель відповідає дійсності на хвостах.

3 МОДИФІКАЦІЯ ОТРИМАНОЇ КОПУЛА-МОДЕЛІ

3.1 Аналіз хвостів розподілу отриманих моделей

Як було зазначено в попередньому розділі, хвости розподілів в досліджуваній моделі мають дуже велике значення. Саме тому, ми зацікавлені провести детальний аналіз хвості отриманої моделі і порівняти з емпіричними даними. На рис. 3.1 зображені суперпозиції емпіричних нижніх (при наближені до 0) та верхніх (при наближені до 1) хвостів до теоретичних, побудованих на основі отриманих моделей. Було вирішено взяти всі 5 тестових моделей для порівняння.

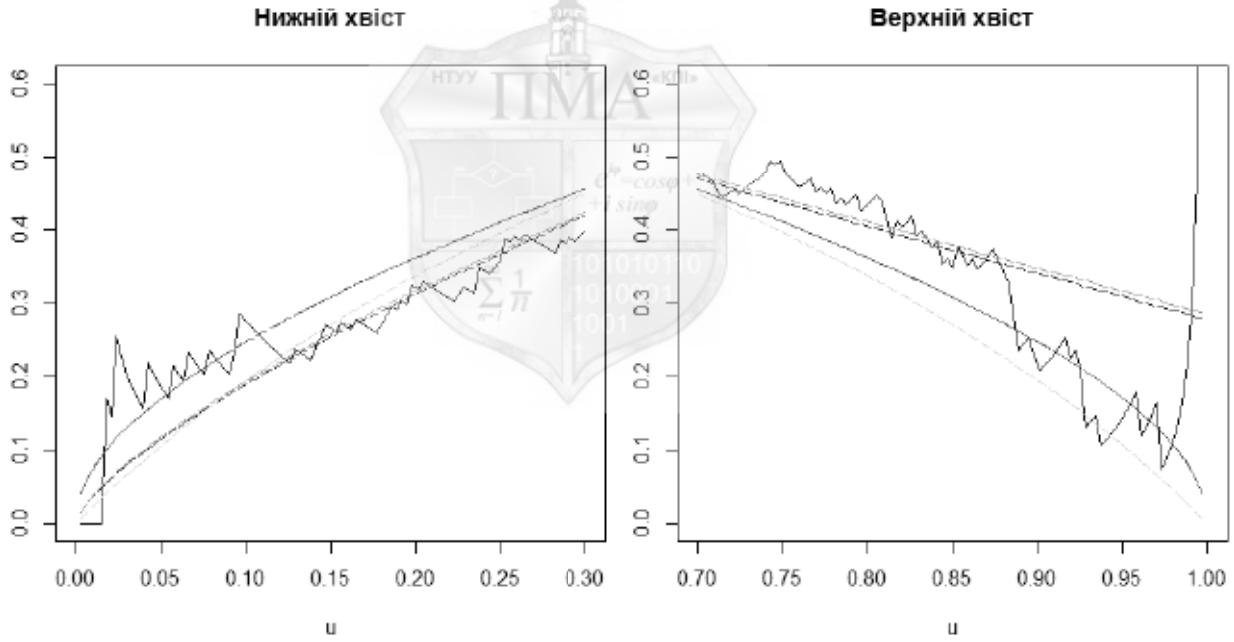

Рисунок 3.1. Аналіз хвостів початкових моделей

На цих малюнках, емпірична модель зображена чорним кольором, Гаусівська копула – фіолетовим, Стьюдента – червоним, Франка – рожевим, Гамбела – синім та Хаслера-Рейса – зеленим.

Одразу слід зазначити, що ситуація на нижньому хвості припустима. Моделі достатньо точно пояснюють залежність. Проте верхній хвіст сильно відрізняється. Легко помітити, що починаючи приблизно з 0,87 значення

емпіричної копули спочатку різко спадають (приблизно до 0,975), а потім роблять стрибок угору, в той час як моделі продовжують повільний спуск. Це означає, що маргінальні розподіли не достатньо точно описують екстремальні значення. Тому необхідно модифікувати маргінальні розподіли за допомогою теореми Пікандса.

3.2 Модифікація маргінальних розподілів

Для модифікації маргінальних розподілів, було вирішено скористатися теоремою Пікандса, або другою теоремою екстремальних значень, що пропонує моделювання важких хвостів за допомогою узагальненого розподілу Парето [21].

Доцільно починати аналіз, з визначення порогів, теореми Пікандаса. Основним і найбільш інтуїтивним методом являється дослідження QQ-графіку (співставлення теоретичних квантилів до емпіричних). При відхиленні від нормалі, можна вважати, що значення перейшли за «поріг» і починається важкий хвіст. В табл. 3.1 наведені результати тесту на відхилення графіку квантилів від нормалі.

Таблиця 3.1 – Результати відхилення від нормалі на QQ-графіку

Випадкова величина	Відхилення знизу (<)	Відхилення зверху (>)
buil	12.53%	80.18%
other	11.35%	88.12%

Візуальну інтерпретацію цього тесту можна спостерігати на рис. 3.2, що добре демонструє відхилення від нормалі і порогові точки.

Таким чином, на визначених верхніх та нижніх хвостах випадкові величини будуть слідувати узагальненим розподілам Парето, а в середині звичайним розподілам Парето, параметри яких вже були оцінені.

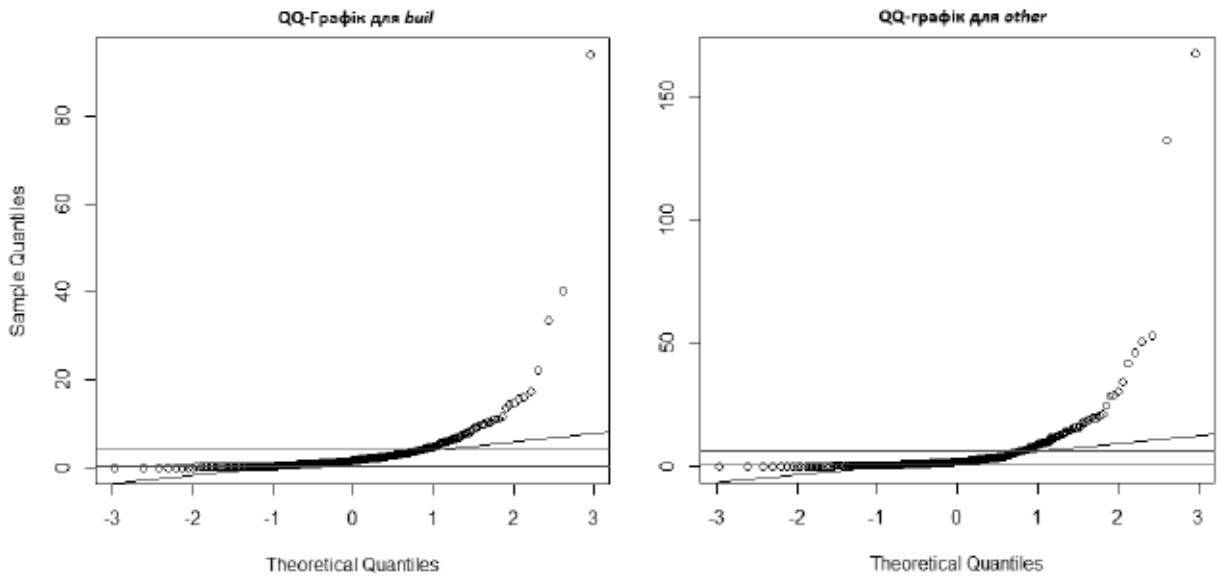

Рисунок 3.2. QQ-графіки для buil та other

За допомогою ММП необхідно оцінити параметри вищезазначених узагальнених розподілів Парето. Слід зазначити, що оцінку необхідно проводити не звичайним ММП, а адаптованим, так як необхідно оцінити параметри для розподілу лише на хвостах. Можливості мови програмування R дозволяють, застосувати метод ММП навіть до складних запрограмованих функцій. Отримані дані оцінки наведені в табл. 3.2.

Таблиця 3.2 – Оцінені параметри для розподілів на хвостах

Випадкова величина	Нижній хвіст		Верхній хвіст	
	μ	σ	μ	σ
buil	0.1511	0.1338	5.3679	10.0416
other	0.1343	0.7655	7.8623	13.1587

Варто зазначити, що так як оцінюються параметри розподілу безпосередньо на хвостах, то параметр локації $u = 0$ автоматично.

З проведених розрахунків випливає, що модифіковані функції розподілу ймовірностей для випадкових величин buil та other приймають наступний вигляд:

– $F_{buil}(x) = \begin{cases} 1 - e^{-\frac{(x-0.1511)}{0.1338}}, x \le 0.2686 \text{ (квантиль } q_{12.53\%}) \\ 1 - \left(\frac{2.6933}{x}\right)^{5.0517}, 0.2686 < x \le 4.0107 \\ 1 - e^{-\frac{(x-5.3679)}{10.0416}}, x > 4.0107 \text{ (квантиль } q_{80.18\%}) \end{cases}$

– $F_{other}(x) = \begin{cases} 1 - e^{-\frac{(x-0.1343)}{0.7655}}, x \le 0.2445 \text{ (квантиль } q_{11.35\%}) \\ 1 - \left(\frac{1.3429}{x}\right)^{2.4800}, 0.2445 < x \le 6.0162 \\ 1 - e^{-\frac{(x-7.8623)}{13.1587}}, x > 6.0162 \text{ (квантиль } q_{88.12\%}) \end{cases}$

Щоб підтвердити ефективність модифікованих маргінальних розподілів, необхідно застосувати КС-тест (результати в табл. 3.3) та побудувати суперпозиції функцій розподілу ймовірностей (рис. 3.3).

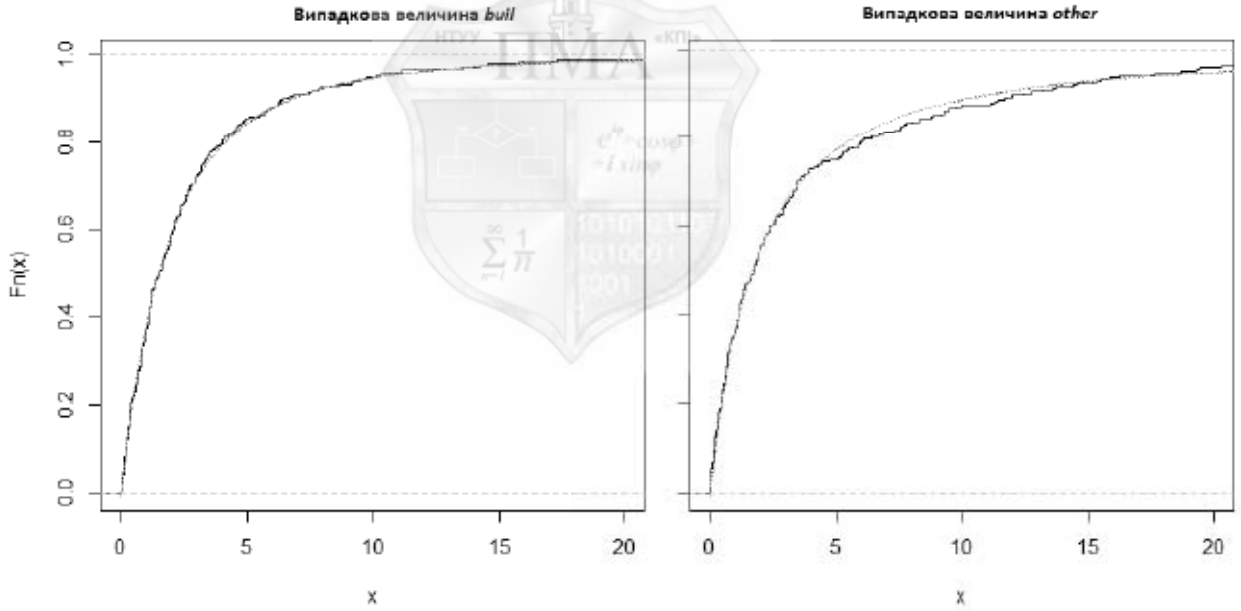

Рисунок 3.3. Суперпозиція ЕФР та двох моделей для buil та other

Таблиця 3.3 – Результати КС-тесту для модифікованих маргіналів

Випадкова величина	D	P-value
buil	0.0158	0.9536
other	0.0212	0.9501

На рис. 3.3 чорним кольором зображені емпіричні моделі, червоним – старі моделі маргінальних розподілів на основі закону Парето, зеленим – модифіковані моделі маргінальних розподілів. Графіки на цьому рисунку можуть мало про що сказати, так як в середині функції ідентичні а відмінності присутні лише на хвостах, проте їх дуже тяжко помітити на такому графіку. Набагато демонстративнішими являються результати КС-тесту. Порівнявши їх з результатами наведеними в табл. 2.2, можна впевнитися, що модифіковані маргінальні розподіли найкраще підходить для моделювання реальних даних. Це вказує на те, що їх можна використати для побудови модифікованої копула моделі.

3.3 Побудова покращеної копули

Так як функції правдоподібності нових маргінальних розподілів являються складними, визначеними на трьох проміжках, то і новий параметр копули буде відрізнятися від попереднього. Оцінка за допомогою ММП дає нове значення параметру копули Хаслера-Рейса з модифікованими маргінальними розподілами для величин buil та other, що дорівнює $\boldsymbol{\alpha = 0.8579}$.

Необхідно перевірити якість отриманої моделі за допомогою коефіцієнту конкордації Кендала. Значення емпіричного тау та теоретичні тау для нової і старої моделі наведені в табл. 3.4.

Таблиця 3.4. Порівняння якості модифікованої та звичайної моделі

Копула	Емпірична	Хаслера-Рейса (з старими маргіналами)	Хаслера-Рейса (з модифікованими маргіналами)
τ (тау)	0.2259	0.2285	0.2253

Модифікована копула-модель, дає кращий результат виходячи з тесту Кендала. Це зумовлено тим, що маргінальні розподіли краще враховують поведінку випадкових величин на хвостах. Графічне відображення значень модифікованої копули на хвостах розподілу, що зображене на рис. 3.4, підтверджує це припущення.

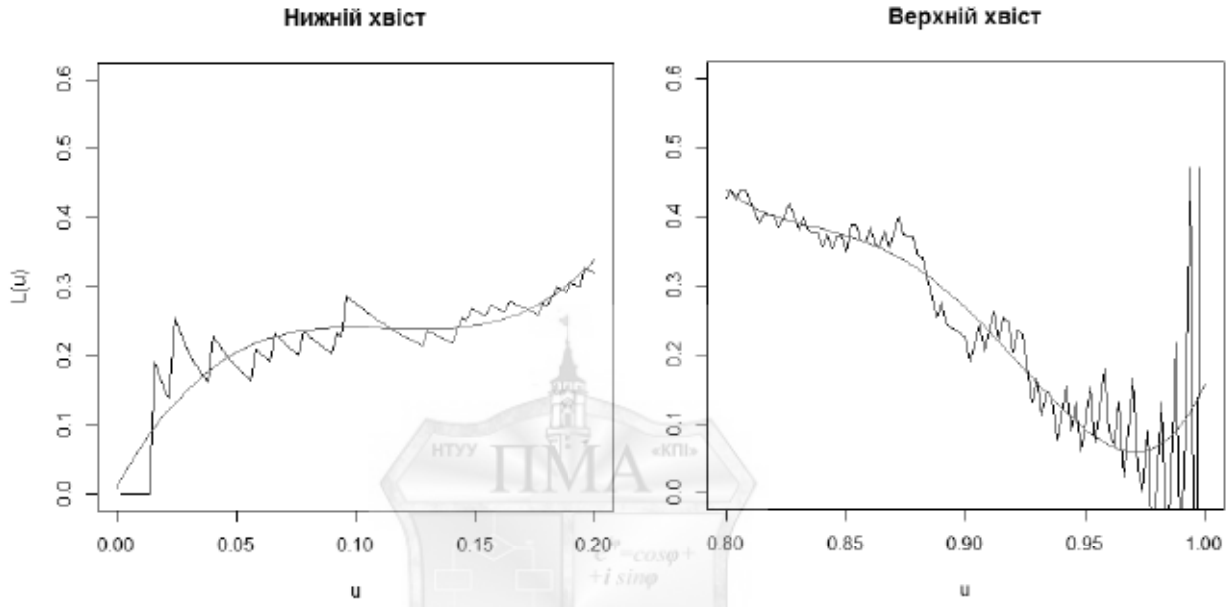

Рисунок 3.4. Поведінка модифікованої моделі на хвостах

Очевидно, що модифікована модель краще адаптована для симуляції реальних страхових даних. Саме тому вона являється результуючою моделлю, з допомогою якої будуть проводитись симуляції та страхові розрахунки.

3.4 Висновки до третього розділу

Результатом цього розділу являється модифікована, остаточна модель залежності збитків нанесених пожежею будівлі та іншим її фінансовим складовим.

Було показано, що врахувавши данні на хвостах розподілу, можна значно покращити якість моделі.

Результуюча копула-модель має наступні параметри :

– Функція розподілу BB buil

$$F_{\text{buil}}(x) = \begin{cases} 1 - e^{-\frac{(x-0.1511)}{0.1338}}, x \le 0.2686 \text{ (квантиль } q_{12.53\%}) \\ 1 - \left(\frac{2.6933}{x}\right)^{5.0517}, 0.2686 < x \le 4.0107 \\ 1 - e^{-\frac{(x-5.3679)}{10.0416}}, x > 4.0107 \text{ (квантиль } q_{80.18\%}) \end{cases}$$

– Функція розподілу BB other

$$F_{\text{other}}(x) = \begin{cases} 1 - e^{-\frac{(x-0.1343)}{0.7655}}, x \le 0.2445 \text{ (квантиль } q_{11.35\%}) \\ 1 - \left(\frac{1.3429}{x}\right)^{2.4800}, 0.2445 < x \le 6.0162 \\ 1 - e^{-\frac{(x-7.8623)}{13.1587}}, x > 6.0162 \text{ (квантиль } q_{88.12\%}) \end{cases}$$

– Залежність між випадковими величинами описується копулою Хаслера-Рейса з параметром $\alpha = 0.8579$.

$$C(X_{\text{buil}}, X_{\text{other}}) = \exp\left(-X_{\text{buil}}\left(1.1656 + 0.429 \ln \frac{\ln X_{\text{buil}}}{\ln X_{\text{other}}}\right) - X_{\text{other}}\left(1.1656 + 0.429 \ln \frac{\ln X_{\text{other}}}{\ln X_{\text{buil}}}\right)\right)$$

4 АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Симуляції на основі отриманої моделі

Тестування статистичних моделей відбувається за допомогою симуляцій та порівняння з реальними даними. Також, основним інтересом для побудови прикладних моделей у страхування являється аналіз властивостей цих моделей та розрахунки на сонові симуляцій

Ідея полягає в тому, щоб відтворити вибірку того ж самого розміру, що і початкова вибірка, взята для побудови моделі. Порівняння основної статистичної інформації може дати непогане уявлення, про якість моделі.

Механізм симуляцій реалізований на основі обернених функцій розподілу та генерації псевдовипадкових величин, що слідують однорідному закону.

Основна статистична інформація довільної симуляції випадкових величин $X \sim F_{buil}$ та $Y \sim F_{other}$ зображені табл. 4.1.

Таблиця 4.1 – Основна статистична інформація симульованих ВВ

Випадкова величина	Кількість спостереж.	Середнє значення	Медіана	Середнє квадрат. відхилення	MIN	MAX	Квантиль 25%	Квантиль 75%
X	327	2.9450	1.6233	6.0051	0.0030	90.1823	0.6353	2.6113
Y	327	5.2188	1.8124	11.8610	0.0103	141.0542	0.8975	3.8265

На цій таблиці добре видно, що середні значення та середньоквадратичні відхилення наближені до даних з табл. 1.2, а мінімальні та максимальні, досить відмінні. Це пов'язано з тим, що як би ми не намагалися підігнати модель на хвостах, вона ніколи не буде ідеальною. Проте і ці результати є досить наближеними.

Варто зазначити, що відтворити точно таку таблицю фактично неможливо, так як результати симуляцій завжди різні.

Щоб отримати більш глибокий аналіз моделі, цікаво буде оглянути асимптотичну статистичну інформацію, тобто провести велику кількість симуляцій та знайти середні оцінки по кожному з параметрів інформації. Згідно з центральною граничною теоремою, чим більше зробимо симуляцій, тим ближче буде результат до реальної статистичної інформації. В табл. 4.2 наведені асимптотичні статистичні дані моделі при $n = 10000$ симуляцій.

Таблиця 4.2 – Асимптотична статистична інформація моделі

Випадкова величина	Кількість спостереж.	Середнє значення	Медіана	Середнє квадрат. відхилення	MIN	MAX	Квантиль 25%	Квантиль 75%
X	327	3.0501	1.5251	6.5812	0.0103	94.1539	0.6201	3.1978
Y	327	5.2487	1.7811	13.6143	0.0012	167.1223	0.5740	4.8788

Не важко побачити, що дані дуже близькі до наведених в таблиці 1.2. Це ще раз підтверджує якість розробленої моделі.

4.2 Розрахунок вартості страхування від пожежі в будівлі

Теорія ймовірностей розроблялася для можливості оцінити та передбачити результати випадково протікаючи процесів. Вона не дозволяє передбачити майбутнє або точно розрахувати скільки потрібно купити лотерейних білетів, щоб зірвати джекпот, проте вона дає уявлення про середні значення та відхилення від них. Саме на цьому побудовані основні засади страхування.

Страхова премія — грошова сума, що її сплачує особа, яка укладає угоду страхування і яка являє собою своєрідну плату за ризик, який бере на себе страхова компанія. Зазвичай страхова премія встановлюється як відсоток від суми угоди страхування, тобто тієї суми, яку страхова компанія сплатить особі у разі настання страхового випадку.

Чиста страхова премія розраховується як:

$$P = T \times C,$$

де T – це тариф, а C – страхова сума.

Нажаль, в нашому випадку, ми маємо дані лише по страхових випадках і за період в 10 років. Найчастіше, контракти, що не підпадають під категорію страхування життя, підписуються не більше ніж на період в 1 рік. Також, для розрахунку тарифу необхідно знати кількість страхових полісів по даному типу ризику. Без цієї інформації неможливо точно розрахувати чисту страхову премію.

Для того, щоб обійти ці проблеми, будуть висунуті деякі гіпотези:

1) Страхова сума дорівнює вартості нещасного випадку, тобто мова йде про повне страхування $C = S$.
2) Страховий контракт підписаний на період в 10 років.
3) Застраховані будівлі в портфелі мають приблизно однакову оціночну вартість.
4) Розмір портфелю становить $N = 1000$ страхових полісів (вибрано довільно, опираючись на логіку).

В такому випадку чиста страхова премія розраховується за формулою:

$$P = \frac{S}{N}$$

Для отримання P, буде використане асимптотичне значення S, що вираховується як середнє від сум страхових випадків великої кількості

симуляцій. Провівши $n = 1000$ отримуємо $\hat{S} = 1184.3561$ млн. данських крон. З цього слідує, що:

$$\widehat{P} = \frac{1184.3561}{1000} = 1.1844 \text{ млн. данських крон}$$

4.3 Висновки до четвертого розділу

В цьому розділі, була перевірена якість симульованих за побудованою моделлю даних. Результати показали, що асимптотичні статистичні дані дуже наближені до реальних, що означає високу точність моделі.

Також, була розрахована чиста страхова премія на основі симульованих даних. Хоча і був взяти фіктивний тип страхового контракту та портфелю, це демонструю область застосування таких копула-моделей. Адже страхові компанії, маючи історичні дані по нещасним випадкам точно будуть мати і інші необхідні дані про портфель.

Також можливе моделювання премій складніших контрактів, де буде використана залежність між випадковими величинами, що буде безпосередньо впливати на розмір премії.

Варто зазначити, що проблемою для такого моделювання, являється розмір вибірки історичних даних, адже в нашому випадку ми мали 327 зафіксованих пожеж на протязі десяти років. Якщо взяти пожежі за один рік, то даних буде мало для побудови гарної моделі. Це є основною проблемою страхото моделювання.

5 ОГЛЯД ПРОГРАМНИХ ЗАСОБІВ

Програмні засоби для побудови та моделювання копула-моделей мають не тривіальну структуру та базуються на складних статистичних обрахунках. Дана програма виконує різноманітні статистичні операції з даними у вигляді векторів та матриць, чисельне оптимізацію та багато інших ітераційних операцій, що потребують точної та швидкої обробки даних. Саме тому, для реалізації системи було обране безкоштовне середовище R, так як воно орієнтоване на статистичні розрахунки і ідеально підходить за своїми функціональними можливостями.

5.1 Вимоги до програмного виробу

Програма повинна отримувати на вхід два вектори з випадковими значеннями однакової довжини. Структура програми передбачає поетапне втручання користувача для аналізу проміжних результатів та прийняття рішень щодо типів розподілів, копул та рівнів значимості. Результатом роботи програми являються оцінені параметри моделі для симуляції залежних випадкових величин. Також існують функції, для безпосереднього використання результуючої моделі для симуляцій та подальших розрахунків.

5.2 Вимоги до системних характеристик

Так як програма написана в середовищі R, то вона може бути запущена лише з цього середовища. Тому основними функціональними

характеристиками програми являються характеристики програмного засобу R (версія не нижча за 3.1.2), а саме:

- ОС: Windows XP, Vista, 7, 8, Linux, MacOS;
- Процесор: будь-який x86 Intel або AMD процесор, що підтримує SSE2 набір команд;
- Простір на диску: 3GB для R, 3-4GB для додаткових пакетів;
- Оперативна пам'ять: 1024MB (рекомендовано 2048MB).
- Враховуючи складність розрахунків, чим потужнішим буде комп'ютер, тим швидше буде працювати програма.

5.3 Вимоги до надійності програмного виробу

Надійність програми полягає в коректній обробці всіх даних і правильній реакції на помилки. Структура програмного засобу передбачає обробку всіх нестандартних ситуацій. Проте, помилки можуть виникати а моменти втручання користувача в наслідок неправильного або не коректного вводу даних. Тому користувач має уважно ознайомитись з інструкціями та вказівками.

5.4 Структура програмного продукту

Програмний продукт представляє собою набір реалізованих функцій і один основний скрипт (.r), що задає порядок викликання цих функцій. Також в програмі використовуються вже готові функції, що підключаються з спеціальних пакетів R. Ці пакети завантажені зі спеціального дзеркала

CRAN, що офіційно розмішує пакети зовнішніх розробників, ретельно перевіривши їх перед цим.

5.4.1 Опис завантажених пакетів R

Як вже зазначалося вище, пакети розроблені різними зовнішніми розробниками. Проте, для того щоб їх розмістити на офіційному дзеркалі CRAN, вони мають відповідати ряду вимог.

Нижче наведено коротке описання основних використаних пакетів [22]:

- fitdistrplus - Розширює функцію fitdistr (пакета MASS) з декількома функціями щоб допомогти підбір параметричного розподілу до цензурованої або не цензурованої інформації. Цензуровані дані можуть містити ліво-цензуровані, прямо-цензуровані та інтервально-цензуровані значення, з деякими нижніми та верхніми межами. На додаток до методу максимальної правдоподібності пакет забезпечує метод моментів, відповідність квантилів та максимальну досконаліть підгонки (доступну тільки для не-цензурованих даних).
- actuar - Додаткова функціонал для реалізації актуарних розрахунків в областях розподілів втрат, теорії ризику (у тому числі теорії розорення), моделювання складних ієрархічних моделей і теорії довіри. Пакет також включає 17 ймовірнісних законів, що зазвичай використовуються в страхуванні, в основному, з важкими хвостами розподілів.
- fCopulae - це набір функцій для управління, дослідження і аналізу двовимірних функцій розподілу. Включені наступні сімейства копул: Архімедові, Еліптичні, Екстремальні та емпіричні.

– gumbel - автономний пакет, що забезпечує функції R для зв'язків Гамбела-Хоугаарда. Надаємо ймовірністні функції (фунції кумулятивного розподілу і функції щільності), функції моделювання (багатовимірна оцінка для копул Гамбела) і функції оцінки (Оцінка максимальної правдоподібності, Оцінка базована на моментах і Канонічна максимальна правдоподібність).

5.4.2 Опис реалізованих функцій

Код нижченаведених функцій наведений в Додатку А.

– remove.naninf – функція, що видаляє з вхідної матриці строчки, що містять дані типу «NaN» та «Inf».
– recap – поєднує в один об’єкт результати оцінки параметрів розподілів та параметру копули.
– fit.cop.IFM – функція, що оцінює параметр копули базуючись на маргінальних розподілах та їх параметрах.
– d(norm)(t)(frank)(gum)(HR)cop – розраховує вектор значень, що відповідають значенням функції щільності відповідної копули.
– tau(norm)(t)(frank)(gum)(HR)cop – розраховує теоретичний коефіцієнт конкордації Кендала, для відповідної копули, опираючись на параметри.
– p(norm)(t)(frank)(gum)(HR)cop – розраховує вектор значень, що відповідають значенням функції розподілу (ймовірностям) відповідної копули.
– Lemp – функція, що будує нижній хвіст емпіричної копули по заданим даним.

- Uemp – функція, що будує верхній хвіст емпіричної копули по заданим даним.
- Lcop – функція, що будує нижній хвіст теоретичної копула-моделі по заданим даним.
- Ucop – функція, що будує нижній хвіст теоретичної копула-моделі по заданим даним.
- optimLQuant – розраховує оптимальне значення квантиля, що відповідає початку нижнього хвосту.
- optimUQuant – розраховує оптимальне значення квантиля, що відповідає початку верхнього хвосту.
- modEstimation – за допомогою ММП оцінює параметри модифікованого розподілу з хвостами.
- modSimul – проводить симуляції згідно з отриманою моделлю.

5.5 Висновки до п'ятого розділу

Розроблені програмні засоби пристосовані для математичних та статистичних розрахунків. Вони адаптовані для моделювання конкретного практичного випадку, проте легко можуть бути використані для роботи з іншими даними.

Скрити R повністю відповідають всім вимогам, дають можливість швидко та точно розрахувати всі необхідні параметри і значення. Програма вимагає втручання користувача на етапах, де необхідно провести аналіз та обрати подальші дії на основі результатів попереднього етапу.

Подальший розвиток даного програмного засобу передбачає розробку інтелектуальних алгоритмів, які зможуть самостійно робити висновки за

результатами кожного етапу і забезпечать отримання результату без втручань користувача.

ВИСНОВКИ

Метою даної магістерської дисертації була розробка алгоритму розрахунку оптимальної копула-моделі для моделювання страхових затрат. Отриманий алгоритм дозволяє швидко і точно побудувати модель на основі заданих даних.

Конкуренція ринку страхових послуг вимагає надточних статистичних розрахунків, чого неможливо досягти без точного моделювання залежностей між випадковими величними. Копула-моделі найкраще адаптовані для пояснення нелінійних залежностей, що найчастіше зустрічаються в реальному житті.

Огляд літератури показав, що існує багато методів та підходів для побудови копула-моделей, проте жодна з них не розрахована на урахування важких хвостів маргінальних розподілів випадкових величин. Саме тому, ключовим етапом алгоритму являється уточнення хвостових розподілів ВВ за допомогою узагальненого розподілу Парето.

Фактично, побудова оптимальної копула-моделі розглядається як наступний алгоритм:

1) Знаходження розподілів ВВ;
2) Оцінка параметрів цих розподілів;
3) Вибір копул кандидатів на пояснення залежності між ВВ;
4) Оцінка параметрів цих копул;
5) Порівняння якості отриманих копула моделей та вибір найкращої;
6) Перевірка поведінки моделі на хвостах;
7) Моделювання хвостів розподілів ВВ за допомогою теореми Пікандса;
8) Оцінка параметра нової копули з модифікованими маргіналами;
9) Перевірка результатів;

Структура розробленої програми являє собою набір запрограмованих функцій, порядок викликання яких задає основний виконавчий скрипт. Для реалізації була обрана мова програмування R, так як вона ідеально підходить для вирішення статистичних задач.

Коректність роботи програми перевірялася на тестовому прикладі, за допомогою симуляцій та огляду асимптотичної статистичної інформації.

Була розроблена модель, що пояснює залежність затрат нанесених пожежею будівлі та іншим її фінансовим складовим.

ПЕРЕЛІК ПОСИЛАНЬ

1. Meneguzzo, David, «Copula sensitivity in collateralized debt obligations and basket default swaps», Journal of Futures Markets T.24, 2003. pp. 37–70.

2. Falissard B., Monga, «Statistique : concepts et méthodes». Paris, Masson, 1993. pp. 14-16.

3. Лоусон Ч., Хенсон Р. «Численное решение задач методом наименьших квадратов», М.: Наука, 1986 – 232 с.

4. Карташов М. В. «Імовірність, процеси, статистика», К.: ВПЦ Київський університет, 2007 – 378 с.

5. Parzen E. «On estimation of a probability density function and mode», Ann. Math. Stat., 1962. pp. 1065-1076.

6. Сеньо П.С. «Теорія ймовірностей та математична статистика», К.: Знання, 2007. – 43 с.

7. Елисеева И. И., Юзбашев М. М. «Общая теория статистики». М: Финансы и Статистика, 2002. – 228–229 сс.

8. Sklar A. «Fonctions de repartition a n dimensions et leurs marges». Publications de l'institut de statistique de l'universite de Paris, 1959. pp. 229–231.

9. Nelsen, Roger B. «An Introduction to Copulas», New York: Springer, 1999. pp. 7–31.

10. Frees E. W., Valdez E. A., «Understanding Relationships Using Copulas», North American Actuarial Journal, 1998. – 2, pp. 1–25

11. Alexander C., Pezier J., «Assessment and Aggregation of Banking Risks», 9th Annual Round Table International Financial Risk Institute (IFCI), 2003. p. 276

12. Kim G., Silvapulle M., Silvapulle P. «Comparison of semiparametric and parametric methods for estimating copulas», Computational Statistics & Data Analysis, 2007. pp. 2836–2850

13. Fisher R.A., Tippett L.H.C. «Limiting forms of the frequency distribution of the largest and smallest member of a sample», Proc. Cambridge Phil. Soc., 1928. pp. 180–190

14. Gnedenko B.V. «Sur la distribution limite du terme maximum d'une serie aleatoire», Annals of Mathematics, 1943. pp. 423–453

15. Pickands J. «Statistical inference using extreme order statistics», Annals of Statistics, 1975. pp. 119–131

16. Лемешко Б.Ю., Помадин С.С., «Проверка гипотез о математических ожиданиях и дисперсиях в задачах метрологии и контроля качества при вероятностных законах, отличающихся от нормального», Метрология. 2004. – 3, 3–15 с.

17. Anderson T. W., Darling D. A. «Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes», Ann. Math. Statist, 1952. V.23, pp. 193–212.

18. Akaike, H. «A new look at the statistical model identification», IEEE Transactions on Automatic Control, 1974. T.19, pp. 716–723.

19. Кобзарь А. И. «Прикладная математическая статистика», М.: Физматлит, 2006. – 624–626 с.

20. Олефір О.С., Савін Є.В. Моделювання залежних збитків від нещасних випадків на основі копула-моделей // Системний аналіз та інформаційні технології: матеріали Міжнародної науково-технічної конференції SAIT 2015, Київ, 22 − 25 червня 2015 р. / ННК «ІПСА» НТУУ «КПІ». — К. : ННК «ІПСА» НТУУ «КПІ», 2015. — С. 108.

21. Олефір О.С., Савін Є.В. Моделювання хвостів розподілів на основі методу ексцесів // Інтелектуальний аналіз інформації: матеріали Міжнародної науково-технічної конференції IAI-2015, Київ, 20 − 22 травня 2015 р. / : сб. тр. – К. : Просвіта, 2015. — С. 158.

22. Інформаційна мережа реєстрації та завантаження пакетів для мови програмування R – CRAN [Електронний ресурс], Режим доступа: http://cran.r-project.org/