

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»**

фізико-математичний факультет

кафедра математичного аналізу та теорії ймовірностей

«На правах рукопису»
УДК _____

До захисту допущено:

Завідувач кафедри

_____ Олег КЛЕСОВ

«14» травня 2021 р.

Магістерська дисертація

на здобуття ступеня магістра

**за освітньо-науковою програмою «Страхова та фінансова
математика»**

зі спеціальності 111 «Математика»

**на тему: «Застосування точкових процесів в узагальненій задачі про
дні народження»**

Виконала:

студентка II курсу магістратури, групи ОМ-91мн
Стаматієва Вікторія В'ячеславівна _____

Науковий керівник:

кандидат фізико-математичних наук, доцент
Ільєнко Андрій Борисович _____

Рецензент:

провідний науковий співробітник відділу функціонального
аналізу інституту математики НАН України, доктор
фізико-математичних наук
Калюжний Олександр Олексійович _____

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць
інших авторів без відповідних
посилань.

Студентка _____

Київ – 2021 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
фізико-математичний факультет

кафедра математичного аналізу та теорії ймовірностей

Рівень вищої освіти – другий (магістерський) за освітньо-науковою програмою

Спеціальність – 111 «Математика»

Освітньо-наукова програма «Страхова та фінансова математика»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Олег КЛЕСОВ

«04» лютого 2021 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Стаматісвій Вікторії В'ячеславівні

1. Тема дисертації «Застосування точкових процесів в узагальненій задачі про дні народження», науковий керівник дисертації Ільєнко Андрій Борисович, кандидат фізико-математичних наук, доцент,

затверджені наказом по університету від «26» березня 2021 р. №901-с

2. Термін подання студентом дисертації 13 травня 2021 року.

3. Об'єкт дослідження послідовність об'єктів довільної природи, що один за одним надходять до спостерігача і кожен з яких незалежно від інших з рівними ймовірностями $\frac{1}{n}$ належить до одного з n класів.

4. Предмет дослідження теореми про збіжність за розподілом вказаних точкових процесів у грубій топології, а також їх наслідки – граничні теореми.

5. Перелік завдань, які потрібно розробити:

- 1) Ознайомитися з вже відомими результатами на тему магістерської дисертації.

- 2) Знайти розподіл перших моментів появ деякого класу в $(r + 1)$ –ий раз та встановити відповідні нормування, спираючись на попередньо доведену граничну теорему для точкових процесів у сенсі т. з. грубої збіжності.
 - 3) Отримати остаточні результати шляхом депуассонізації.
 - 4) Візуалізувати дані в середовищі RStudio.
 - 5) Розглянути різноманітні застосування цих результатів.
6. Орієнтовний перелік графічного (ілюстративного) матеріалу 27 слайдів.
7. Орієнтовний перелік публікацій
- 1) В. В. Стаматієва. Застосування точкових процесів в узагальненій задачі про дні народження. // X Всеукраїнська наукова конференція молодих математиків. Нац. техн. ун-т України «КПІ ім. І. Сікорського». – Київ. – 2021. – с.42.
8. Дата видачі завдання 04 лютого 2021 року.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Отримання та обговорення основної задачі дипломної роботи	04.02.21 – 07.02.21	виконано
2.	Огляд літератури	08.02.21- 18.02.21	виконано
3.	Дослідження спорідненої задачі про збирача купонів	19.02.21 – 09.03.21	виконано
4.	Доведення допоміжної леми про збіжність точкового процесу до неоднорідного процесу Пуассона в пуассонізованій постановці	10.03.21 – 15.03.21	виконано

5.	Доведення основної теореми шляхом депуассонізації	16.03.21 – 10.04.21	виконано
6.	Огляд різноманітних застосувань одержаних результатів	11.04.21 – 20.04.21	виконано
7.	Візуалізація даних шляхом імітаційного моделювання в середовищі RStudio	21.04.21 – 28.04.21	виконано
8.	Оформлення магістерської дисертації	29.04.21 – 07.05.21	виконано

Студент

Вікторія СТАМАТІЄВА

Науковий керівник

Андрій ІЛЬЄНКО

Реферат

Магістерська дисертація: 60 сторінок, 27 слайдів для проектора, 17 першоджерел, 11 ілюстрацій.

Актуальність магістерської дисертації обумовлюється численними застосуваннями різних варіантів задачі про дні народження в багатьох предметних областях (інформатика, криптографія, генетика тощо), а також сучасною тенденцією дослідження асимптотичних властивостей дискретних імовірнісних моделей за допомогою методів теорії точкових процесів.

Метою роботи є дослідження та застосування збіжності точкових процесів, пов'язаних з узагальненою задачею про дні народження.

Об'єктом дослідження є послідовність об'єктів довільної природи, що один за одним надходять до спостерігача і кожен з яких незалежно від інших з рівними ймовірностями $\frac{1}{n}$ належить до одного з n класів. У класичній задачі про дні народження об'єкти асоціюються з людьми, а класи об'єднують людей зі спільними днями народження.

Предметом дослідження є теореми про збіжність за розподілом вказаних точкових процесів у грубій топології, а також їх наслідки — граничні теореми для різних змістовних функціоналів від таких процесів.

Для доведення цих результатів використовувався апарат теорії ймовірностей та теорії точкових процесів (зокрема, методи пуассонізації та депуассонізації). Крім того, було проведено імітаційне моделювання в середовищі RStudio, що дало змогу перевірити та візуалізувати отримані твердження.

Центральним результатом роботи є гранична теорема про збіжність точкових процесів, що описують моменти надходження $(r + 1)$ –их представників різних класів, до деякого неоднорідного точкового процесу Пуассона. Зокрема, в роботі знайдено міру інтенсивності цього граничного процесу та вигляд необхідного нормування. Наслідками цього твердження є граничні теореми для різноманітних числових характеристик моделі.

Результати роботи доповідалися на X Всеукраїнській науковій конференції молодих математиків.

Ключові слова: задача збирача купонів, узагальнена задача про дні народження, точковий процес, пуассонізація, депуассонізація, неоднорідний процес Пуассона, «груба» збіжність.

Abstract

Master degree thesis contains 60 pages, 27 slides for projector, 17 primary sources, 11 illustrations.

The relevance of the master's dissertation is determined by the numerous applications of different variants of the birthdays problem in many areas (computer science, cryptography, genetics, etc.), as well as by the current trend of studying the asymptotic properties of discrete probability models applying the point process theory.

The aim of the work is to study and apply the convergence of point processes associated with the generalized birthday problem.

The object of study is a sequence of objects of arbitrary nature, one after another meeting the observer and each of which, independently of the others with equal probabilities $\frac{1}{n}$ belongs to one of the n classes. In the classic birthday problem, objects are associated with people, and classes unite people with shared birthdays.

The subject of the research is the theorems on the convergence of the distribution of these point processes in a vague topology, as well as their consequences — limit theorems for different content functionals from such processes.

To come to these results, we used the apparatus of probability theory and the theory of point processes (in particular, the methods of Poissonization and depoissonization). In addition, simulation was performed in the RStudio environment, which allowed to verify and visualize the obtained statements.

The central result of the work is the limit theorem on the convergence of point processes, which describe the moments of receipt of $(r + 1)$

representatives of different classes to some non-homogeneous Poisson point process. In particular, the intensity measure of this limiting process and the form of the necessary rationing are found in the work. The consequences of this statement are limit theorems for various numerical characteristics of the model.

The results were presented at the X All-Ukrainian Scientific Conference of the Young Mathematicians.

Key words: the coupon collector's problem, generalized birthday problem, point process, poissonization, depoissonization, non-homogeneous Poisson point process, «vague» convergence.

Зміст

Вступ	10
Огляд задачі про збирача купонів	14
1.1. Теоретичні відомості.....	14
1.2. Застосування точкових процесів у задачі про збирача купонів...	18
Узагальнена задача про дні народження.....	23
2.1. Постановка задачі	23
2.2. Позначення та зауваження.....	24
2.3. Основний результат	27
Застосування основного результату	34
Візуалізація даних в середовищі RStudio	43
4.1. Візуалізація I	43
4.2. Візуалізація II.....	51
4.3. Візуалізація III	55
Висновки	58
Список використаних джерел	59

Вступ

Задача про дні народження слугує яскравим прикладом задач комбінаторної теорії ймовірностей. Наведемо одне з її численних формулювань у вигляді, зручному для подальших узагальнень.

Розглянемо достатньо велику множину людей, які в послідовні цілі моменти часу один за одним заходять до кімнати. Позначимо через $T_1^{(365)}$ (випадковий) момент часу, коли до кімнати вперше зайшла людина, день народження якої збігається з днем народження когось з вже присутніх. Тут верхній індекс показує загальну кількість різних можливих днів народження (всі роки вважаються невисокосними), а нижній означає, що достатньо збігу з днем народження лише *одного* з присутніх.

Парадокс цієї задачі полягає в тому, що ймовірність зустрічі людей з однаковим днем народження залежить не лише від кількості цих осіб, а й від кількості самих збігів.

В даній роботі ми узагальнимо задачу про дні народження на випадок довільної кількості класів (наприклад, $n = 365$ днів) та фіксованої кількості збігів. Позначимо через величину $T_r^{(n)}$ перший момент часу, коли деякий з класів з'явився в $(r + 1)$ -ий раз.

У випадку $r = 1$ математичне сподівання величини $T_1^{(n)}$ можна обчислити за наступними формулами (див.[1]):

$$\mathbb{E}T_1^{(n)} = 1 + \sum_{k=1}^n \frac{n(n-1) \dots (n-k+1)}{n^k},$$

$$\mathbb{E}T_1^{(n)} = \int_0^{\infty} e^{-t} \left(1 + \frac{t}{n}\right)^n dt.$$

Перевагою останньої формули над попередньою є те, що представлення в інтегральній формі більш природно використовувати у багатьох комбінаторних умовах.

Спрямувавши $n \rightarrow \infty$, можна показати, що

$$\mathbb{E}T_1^{(n)} = \sqrt{\frac{n\pi}{2}} + \frac{2}{3} + O\left(n^{-\frac{1}{2}}\right),$$

див. [1].

При $n = 365$ математичне сподівання величини $T_1^{(n)}$ наближено дорівнює 24.6, що трохи відрізняється від відомого медіанного значення 23.

У книзі Ф. Флажолє та Р. Седжвіка [1] було наведено наступну формулу для обчислення очікуваного часу, необхідного для першої появи події « $(r + 1)$ осіб мають однакові дні народження»:

$$I(n, r) := \int_0^\infty e^{-t} e_r\left(\frac{t}{n}\right)^n dt,$$

$$\text{де } e_r(z) := 1 + \frac{z}{1!} + \frac{z^2}{2!} + \dots + \frac{z^r}{r!}.$$

(При $r = 1$ та $n = 365$ одержуємо звичайну задачу про дні народження).

Асимптотику для $I(n, r)$ вперше було одержано Ньюманом та Кламкіном в 1967 році (див. [14]). Вона визначається наступним чином:

$$I(n, r) \underset{n \rightarrow \infty}{\sim} r^{+1} \sqrt{(r+1)!} \Gamma\left(1 + \frac{1}{r+1}\right) n^{1-\frac{1}{r+1}}.$$

Наведемо деяку реалізацію задачі про дні народження на рис. 1 (див. [1]):

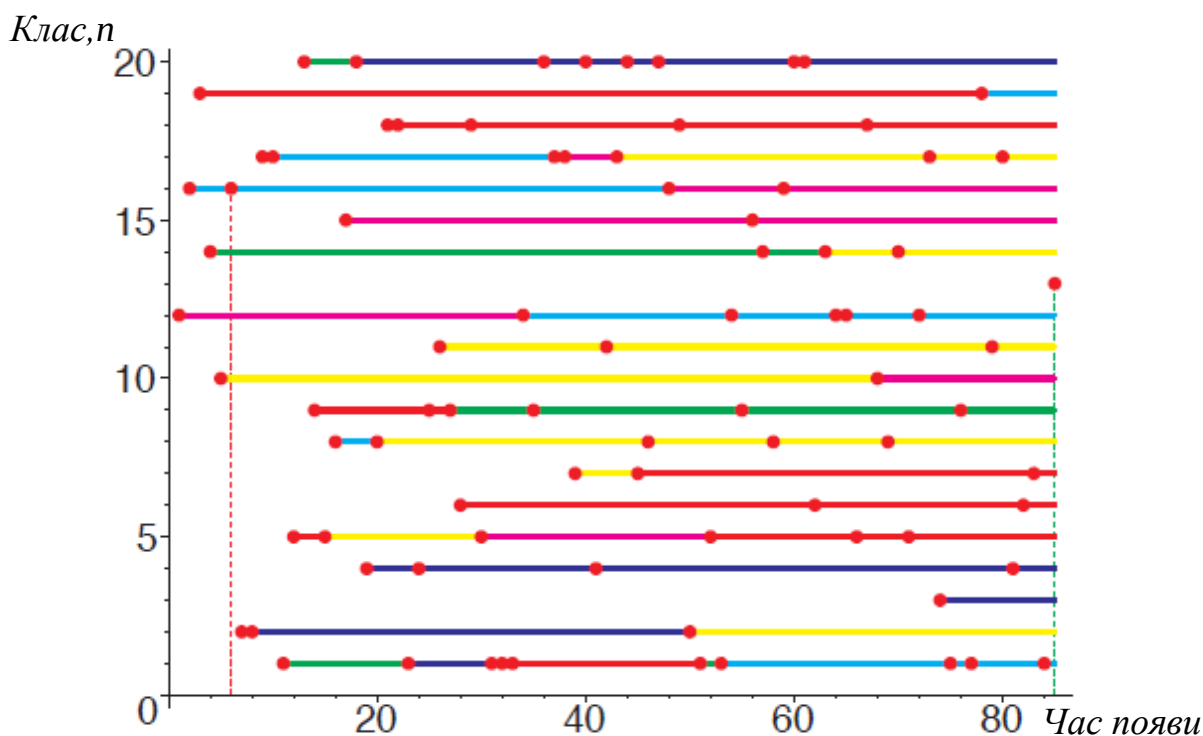


Рис. 1.

З наших результатів випливає, що точкові процеси, які описують появи $(r + 1)$ -их представників з різних n класів, $r \geq 1$, (тобто червоних точок у кожному рядку), збігаються при $n \rightarrow \infty$ до неоднорідного процесу Пуассона з мірою інтенсивності

$$\lambda_r(dx) = \frac{x^r}{r!} \cdot \mathbb{I}\{x \geq 0\}dx.$$

Дана задача є досить актуальною. Вона має безліч застосувань у різних галузях. Наведемо кілька прикладів:

- Дослідження збігів ДНК

Ймовірність того, що у двох випадково обраних людей зі всієї Земної кулі будуть ідентичні ДНК, рівна $\frac{1}{10^{12}}$ (звичайно, ми виключаємо однойцевих близнюків). Якщо ж провести подібний експеримент, але при цьому обмежити площу, скажімо, обрати лише для зіставлення ДНК-

маркерів двох американців, то ймовірність суттєво зміниться. Вірогідність знайти двох «генетичних близнюків» серед 328 мільйонів американців становитиме $1/3000$.

Зі статистики кримінальних злочинів відомо, що кожного року в США стається приблизно 16 000 вбивств, тоді, відповідно до вищеподаної ймовірності, ДНК вбивць в 5 випадках з 16000 будуть однаковими з ДНК громадян Америки, які непричетні до злочинів.

- Космічне сміття

Ймовірність того, що відбудеться зіткнення двох орбітальних об'єктів – нескінченно мала, але що якщо ми розглянемо зіткнення супутника з певним об'єктом, діаметр якого є більшим за 1 см? Звичайно, що ймовірність зіткнення значно збільшується, адже таких об'єктів понад 900 000. Тому якщо ухилитись від супутника легко, і це можливо контролювати, то зі зростанням чисельності космічного сміття ухилитись від будь-яких інших зіткнень стає дедалі важчим.

- Хеш-функції

Яскравим прикладом парадокса днів народження є використання хеш-функцій. Припустимо, є Боб, який підписує контракти в цифровій формі, а у нас є два документи: справжній та підроблений (скажімо, змінене форматування відносно справжнього). Нам же потрібно, аби він пропустив фальшивий документ, а це можливо лише за умови, що хеш-значення обох документів ідентичні, чи це можливо взагалі? Відповідь: так. Парадокс днів народження говорить про те, що ймовірність знайти однакові хеш-значення для документів є значною, тому, виходячи з цього факту, захист інформації трансформується кожен день.

Розділ 1

Огляд задачі про збирача купонів

1.1. Теоретичні відомості

Розглянемо споріднену задачу до нашої, а саме задачу про збирача купонів. Вона є типовим прикладом задач комбінаторної теорії ймовірностей. Дану задачу можна сформулювати наступним чином: людина збирає купони, кожен з яких належить до одного з n різних типів. Купони надходять один за одним в різний час, тип кожного купона є рівноймовірним і незалежним від типу попереднього.

Історія задачі збирача купонів почалася в 1708 році, коли дане питання вперше з'явилося у *De Mensura Sortis*, написаною А. де Муавром. Більше результатів, в роботах Ейлера та Лапласа, було отримано у випадку рівних ймовірностей, тобто при $p_k \equiv \frac{1}{n}$ для будь-якого k .

У 1954 р. Х. фон Шеллінг вперше одержав результати часу очікування для зібрання колекції у випадку, коли ймовірність прибуття кожного купона не була сталою і в 1960 р. Д. Ньюмен та Л. Шепп підраховали середній час очікування для зібрання двох купонів у випадку рівних ймовірностей.

Задача збирача купонів має багато застосувань: в електротехніці, а саме проблеми помилок кеша; в біології, де вона використовується для оцінки кількості видів тварин, і т.д.

Припустимо, що існує N різних купонів з рівними ймовірностями, з ймовірністю придбання будь-якого типу в будь-який час, що дорівнює $\frac{1}{N}$. Надалі наведемо кілька підходів до знаходження очікуваної кількості купонів, які потрібно придбати для збору колекції (див. [2]).

Геометричний закон розподілу

Нехай X – кількість купонів, де X – випадкова величина. Ми можемо представити $X = X_1 + \dots + X_N$, де для будь-якого $i = 1, 2, \dots, N$, X_i – кількість купонів, які нам потрібно придбати, щоб перейти від $(i - 1)$ до i різних типів купонів. Нехай $X_1 = 1$, оскільки ми розглядаємо рівномірний розподіл, то звідси випливає, що коли i різних купонів зібрані, то новий купон, який буде куплено, буде іншого типу з ймовірністю, рівною $\frac{N-i}{N}$. За нашим припущенням про незалежність, ми отримуємо, що випадкова величина X_i , для будь-якого $i \in \{2, \dots, N\}$, незалежна від інших випадкових величин і має геометричний розподіл з параметром $\frac{N-i+1}{N}$. Очікувана кількість купонів, які ми купуємо для повного зібрання, рівна

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X_1] + \dots + \mathbb{E}[X_N] = \\ &= 1 + \frac{N}{N-1} + \frac{N}{N-2} + \dots + \frac{N}{2} + N = N \sum_{i=1}^N \frac{1}{i}. \end{aligned}$$

Підхід за допомогою ланцюгів Маркова

Навіть якщо попередній результат дуже простий і формула визначена, надалі буде запроваджено альтернативний підхід, який вдасться використовувати у випадку, коли ситуація стає більш складною. Якщо припустити, що один купон надходить у будь-яку одиницю часу, ми можемо інтерпретувати попередні величини X_i , як додатковий час, який нам доведеться зачекати, щоб отримати i -ий купон після $(i - 1)$ різних типів купонів, що були одержані. Дану задачу можна розв'язати за допомогою ланцюгів Маркова.

Нехай Y_n – кількість купонів різних типів, які були одержані після n одиниць часу і, виходячи з того, що ймовірність знайти купон будь-якого

типу у будь-який момент часу рівна $p = \frac{1}{N}$, Y_n будуть ланцюгом Маркова на $S = \{0, 1, \dots, N\}$, де $|S| = N + 1$, це безпосередньо означає, що матриця переходу визначається як

$$P = \begin{pmatrix} 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \frac{1}{N} & \frac{N-1}{N} & 0 & \dots & \dots & 0 \\ 0 & 0 & \frac{2}{N} & \frac{N-2}{N} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & \frac{N-1}{N} & \frac{1}{N} \\ 0 & \dots & \dots & \dots & 0 & 0 & 1 \end{pmatrix}.$$

Зауважимо, що $\{N\}$ – це замкнений клас, і стан N є поглинаючим. Отже, для визначення середнього часу, необхідного для досягнення стану N зі стану X_n , який дорівнює очікуваній кількості купонів, потрібних для придбання, потрібно розв’язати лінійну систему

$$\begin{cases} k_N = 0, \\ k_i = 1 + \sum_{j \neq N} p_{ij} k_j, i \neq N. \end{cases}$$

Використовуючи матрицю переходу P , ця система еквівалентна наступній:

$$\begin{cases} k_N = 0, \\ k_{N-1} = N, \\ k_{N-2} = k_{N-1} + \frac{N}{2} = N + \frac{N}{2}, \\ k_{N-3} = k_{N-2} + \frac{N}{3} = N + \frac{N}{2} + \frac{N}{3}, \\ \dots \\ k_2 = k_3 + \frac{N}{N-2} = N + \frac{N}{2} + \dots + \frac{N}{N-3} + \frac{N}{N-2}, \\ k_1 = k_2 + \frac{N}{N-1} = N + \frac{N}{2} + \dots + \frac{N}{N-2} + \frac{N}{N-1}, \\ k_0 = k_1 + 1 = N + \frac{N}{2} + \dots + \frac{N}{N-2} + \frac{N}{N-1} + 1. \end{cases}$$

Звідси випливає, що час очікування для зібрання всіх N купонів задається такою формулою:

$$k_0 = N \sum_{i=1}^N \frac{1}{i}.$$

Наближені формули

Для малих значень N ми можемо легко обчислити очікувану кількість купонів, які потрібно придбати для формування повної колекції:

N	1	2	3	4	5	6	7
$\mathbb{E}[X]$	1	3	5.5	8.33	11.42	14.7	18.15

Для великих значень N ми можемо використати наступні наближення:

$$\sum_{i=1}^N \frac{1}{i} = \log(N) + \gamma + \frac{1}{2N} + O\left(\frac{1}{N^2}\right),$$

де $\gamma \approx 0,5772156649$ – стала Ейлера-Маскероні. Ми можемо приблизно обчислити $\mathbb{E}[X]$ при $N \rightarrow +\infty$ наступним чином:

$$\mathbb{E}[X] = N \log(N) + N\gamma + \frac{1}{2} + O\left(\frac{1}{N}\right), N \rightarrow +\infty.$$

При $N = 100$ точна оцінка кількості купонів рівна $\mathbb{E}[X] = 518.7377 \pm 10^{-4}$, в той час як оцінка $N \log(N) + N\gamma + \frac{1}{2}$ при $N = 100$ рівна $\mathbb{E}[X] = 518.7385 \pm 10^{-4}$, як наближене значення $\mathbb{E}[X]$.

1.2. Застосування точкових процесів у задачі про збирача купонів

Нехай $T_c^{(n)}$ — позиція (випадкової) кількості купонів, які людина повинна зібрати для одержання $c \in \mathbb{N}$ повних колекцій. В наш час є багато відомих результатів, пов'язаних з асимптотикою $\mathbb{E}T_c^{(n)}$, а також з граничною поведінкою величин $T_c^{(n)}$. Одним з класичних та добре відомих результатів є

$$\mathbb{E}T_c^{(n)} = n \ln n + (c-1)n \ln \ln n + (\gamma - \ln(c-1)!)n + \mathcal{O}(n),$$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{T_c^{(n)}}{n} - \ln n - (c-1) \ln \ln n < x \right\} = \exp \left\{ -\frac{e^{-x}}{(c-1)!} \right\},$$

де $\gamma = -\Gamma'(1)$ — стала Ейлера-Маскероні (див., напр., [4]).

Широко розвинутою ідея пуассонізації була у роботі Ільєнка А. Б. [3], завдяки якій вдалося отримати деякі узагальнення та нескінченновимірні розширення класичних граничних теорем. Зокрема, було доведено наступну теорему (див. теорему 3.1 в [3]):

Теорема 1.1. *Нехай $Y_{i,r}^{(n)}, i \in \mathbb{N}_n = \{1, \dots, n\}, r, n \in \mathbb{N}$, — момент надходження r — того купона типу i , δ_u — точкова міра Дірака, тобто $\mathbb{I}\{u \in \cdot\}$. Уведемо на $\mathbb{R}$ точковий процес*

$$\xi_r^{(n)} = \sum_{i=1}^n \delta_{\frac{Y_{i,r}^{(n)}}{n} - \ln n - (r-1) \ln \ln n},$$

а через ξ_r позначимо точковий процес Пуассона з мірою інтенсивності

$$\lambda_r(dx) = \frac{1}{(r-1)!} e^{-x} dx.$$

Тоді $\xi_r^{(n)} \xrightarrow{dv} \xi_r$ при $n \rightarrow \infty$.

Граничний процес ξ_r має просту інтерпретацію. Нехай ξ є стаціонарним точковим процесом Пуассона з одиничною інтенсивністю, обмеженим на $(0, +\infty)$, і покладемо

$$h(x) = -\ln(r-1)! - \ln x, \quad x > 0.$$

Тоді $\xi_r = \xi \circ h^{-1}$ за розподілом. Інакше кажучи, $\xi_r = \sum_{x \in \text{supp} \xi} \delta_{h(x)}$ за розподілом.

Дійсно, за теоремою про відображення (див. теорему 5.1 в [5]), $\xi \circ h^{-1}$ є пуассонівським процесом з мірою інтенсивності вигляду $\text{Leb} \circ h^{-1}$, де Leb означає міру Лебега. Тоді, для будь-яких $[a, b] \subset \mathbb{R}$,

$$(\text{Leb} \circ h^{-1})[a, b] = \text{Leb} \left[\frac{e^{-b}}{(r-1)!}, \frac{e^{-a}}{(r-1)!} \right] = \int_a^b \frac{1}{(r-1)!} e^{-x} dx = \lambda_r[a, b],$$

що й доводить результат.

З вищенаведеної теореми випливає ряд відомих граничних результатів, які спочатку доводилися прямими складними розрахунками. Більше того, цей підхід дозволяє отримати вагомі узагальнення та нескінченновимірні розширення цих результатів.

Як приклад, розглянемо $T_{r,m}^{(n)}$, $m \leq n-1$, моменти часу, коли в системі деякі $(n-m)$ з n типів купонів вже з'явилися в системі r разів кожен, і поставимо $T_{r,m}^{(n)} = 0$ для $m \geq n$. У [6], [13], [9] вивчалися різні граничні теореми для випадку $r = 1$. Зокрема, теорема 4 у [6] стверджує, що

$$\ln 2n - \frac{T_{1,m}^{(n)}}{n} \xrightarrow{d} \ln Q, \quad n \rightarrow \infty,$$

де Q належить χ_{2m+2}^2 , де χ^2 – розподіл з $(2m+2)$ ступенями свободи. Теорема 4.1 з [3], є узагальненням для довільного $r \in \mathbb{N}$ та розширенням в сенсі збіжності за розподілом в $\mathbb{R}^\infty$, а також пояснює появу χ^2 – розподілу. Сформулюємо дану теорему:

Теорема 1.2. Нехай задано $V_r^{(n)}$ та V_r в $\mathbb{R}^\infty$, де

$$V_r^{(n)} = \left(\frac{T_{r,m}^{(n)}}{n} - \ln n - (r-1) \ln \ln n, m \in \mathbb{N} \cup \{0\} \right),$$

$$V_r = \left(-\ln(r-1)! - \ln \sum_{j=1}^{m+1} E_j, m \in \mathbb{N} \cup \{0\} \right),$$

а $E_j, j \in \mathbb{N}$, н.о.р.в.в. з $\text{Exp}(1)$.

Тоді виконується $V_r^{(n)} \xrightarrow{d} V_r$ в $\mathbb{R}^\infty$ при $n \rightarrow \infty$.

Зауваження 1.1. Зосередившись на одновимірному випадку, ми отримуємо

$$\frac{T_{r,m}^{(n)}}{n} - \ln n - (r-1) \ln \ln n \xrightarrow{d} -\ln(r-1)! - \ln S_{m+1}, \quad n \rightarrow \infty,$$

де $S_{m+1} = \sum_{j=1}^{m+1} E_j \sim \Gamma(m+1, 1)$.

Зупинившись на випадку $r = 1$, ми дістанемо

$$\ln 2n - \frac{T_{1,m}^{(n)}}{n} \xrightarrow{d} \ln(2S_{m+1}), \quad n \rightarrow \infty.$$

Остаточно, зауважимо, що $2S_{m+1} \sim \chi_{2m+2}^2$.

В роботі [3] розглянуто ще одне застосування теореми 1.1. Воно пов'язане з «рідкісними» типами купонів. Нагадаємо, що купон, що надходить в систему, належить до будь-якого з n типів з однаковою ймовірністю $\frac{1}{n}$. Але, через випадкові фактори, деяким типам купонів знадобиться тривалий час, поки вони не надійдуть в r – тий раз. Назвемо тип $i \in \mathbb{N}_n$ x -рідкісним, $x \in \mathbb{R}$, якщо

$$Y_{i,r}^{(n)} \geq nx + n \ln n + (r-1)n \ln \ln n.$$

Позначимо через $C_r^{(n)}(x)$ кількість x -рідкісних типів:

$$C_r^{(n)}(x) = \sum_{i=1}^n \mathbb{I} \{ Y_{i,r}^{(n)} \geq nx + n \ln n + (r-1)n \ln \ln n \}.$$

Нижче ми сформулюємо функціональну граничну теорему для $C_r^{(n)} = (C_r^{(n)}(x), x \in \mathbb{R})$ у J_1 –топології Скорохода. Її детальне доведення можна знайти у [3].

Нехай $N = (N(t), t \geq 0)$ – однорідний процес Пуассона з одиничною інтенсивністю, розглядається не випадкова точкова міра, а класична, як процес Леві з приростом Пуассона. Власне, $N(t) = \xi(0, t]$, де ξ є стаціонарним точковим процесом Пуассона з одиничною інтенсивністю, обмеженим на $(0, +\infty)$. Також нехай $N_r(x) = N\left(\frac{e^{-x}}{(r-1)!}\right)$, $x \in \mathbb{R}$.

Теорема 1.3. *Справедливе наступне: $C_r^{(n)} \xrightarrow{d} N_r$ в $D(\mathbb{R})$ з J_1 –топологією, при $n \rightarrow \infty$.*

Зауваження 1.2. *Строго кажучи, процеси $C_r^{(n)}$ і N_r є càglàd, але не càdlàg, тому ми маємо розглянути $D_{\text{left}}(\mathbb{R})$, простір неперервних зліва функцій з правосторонніми границями, замість $D(\mathbb{R})$. Але оскільки J_1 -*

топология може бути введена на $D_{left}(\mathbb{R})$ так само, як і на $D(\mathbb{R})$, ми закриємо очі на ці відмінності (див. зауваження 3.2 на с. 58 в [8]).

Розділ 2

Узагальнена задача про дні народження

2.1. Постановка задачі

Узагальнену задачу про дні народження можна сформулювати в такий спосіб. Розглянемо нескінченну послідовність об'єктів (наприклад, людей), кожний з яких незалежно від інших з однаковими імовірностями $\frac{1}{n}$ належить до одного з n класів (наприклад, народжений в один з $n = 365$ днів року). Об'єкти надходять один за одним у послідовні цілі моменти часу.

Для фіксованого $r \geq 1$ розглянемо випадкову величину $T_r^{(n)}$, яка задає перший момент надходження об'єкту, клас якого вже зустрічався рівно r разів. Будемо розглядати ситуацію, коли r – фіксоване, а $n \rightarrow \infty$.

Ми цікавимося асимптотичною поведінкою величин $T_r^{(n)}$, або, більш точно, встановленням їх граничних розподілів при $n \rightarrow \infty$ та відповідних нормувань.

Таку асимптотичну поведінку можна встановити, якщо для різних n ввести точкові процеси моментів $(r + 1)$ -их появ різних класів та дослідити граничний розподіл перших (найменших) точок цих процесів. Цей розподіл можна знайти, спираючись на попередньо доведену граничну теорему для самих точкових процесів у сенсі т. з. грубої збіжності.

2.2. Позначення та зауваження

Нехай $Y_{i,r}^{(n)}, i \in \mathbb{N}_n = \{1, \dots, n\}, r, n \in \mathbb{N}$, — момент надходження $(r + 1)$ -го об'єкту класу i . Тоді $Y_{i,r}^{(n)} \sim \text{NegBin}\left(r + 1, \frac{1}{n}\right)$, де NegBin — це від'ємний біноміальний розподіл кількості випробувань до появи $(r + 1)$ -го успіху:

$$\mathbb{P}\{Y_{i,r}^{(n)} = k\} = \binom{k-1}{r} \left(\frac{1}{n}\right)^{r+1} \left(1 - \frac{1}{n}\right)^{k-r-1}, k \geq r + 1.$$

Для фіксованих значень r та n , випадкові величини $Y_{i,r}^{(n)}, i \in \mathbb{N}_n$, є однаково розподіленими, але не незалежними, оскільки $Y_{i,r}^{(n)} \neq Y_{i',r}^{(n)}$, при $i \neq i'$.

Із залежними випадковими величинами працювати складно, тому ми введемо процес так званої пуассонізації, щоб позбутися цієї залежності.

Припустимо, що об'єкти надходять у випадкові моменти часу з незалежними $\text{Exp}(1)$ — розподіленими випадковими величинами $E_j, j \in \mathbb{N}$. Більш формально, ми вводимо одинично-розподілений незалежний пуассонівський точковий процес $\Xi^{(n)} = \sum_{k=1}^{\infty} \delta_{(X_k, M_k)}$ з рівномірним на $\mathbb{N}_n$ розподілом: $\mathbb{P}\{M_k = i\} = \frac{1}{n}, i \in \mathbb{N}_n$. Тут $X_k = \sum_{j=1}^k E_j$ — час прибуття, M_k — типи об'єктів, що надходять у систему, а δ_u — міра Дірака, тобто $\mathbb{I}\{u \in \cdot\}$. Тоді (див. теорему 5.8 з [5])

$$\Xi_i^{(n)} = \sum_{\substack{(X_k, M_k) \in \\ \text{supp} \Xi^{(n)}}} \delta_{X_k} \mathbb{I}\{M_k = i\} = \Xi^{(n)}(\cdot \times \{i\}), i \in \mathbb{N}_n,$$

є незалежним $\frac{1}{n}$ — розподіленим пуассонівським точковим процесом. Процес $\Xi_i^{(n)}$ описує прибуття об'єкту i — того класу. Користуючись цими

позначеннями, вищевказана випадкова величина $Y_{i,r}^{(n)}$ має наступне представлення:

$$Y_{i,r}^{(n)} = \min \left\{ m \geq r + 1 : \sum_{k=1}^m \mathbb{I}\{M_k = i\} = r + 1 \right\}. \quad (2.1)$$

Нехай $Z_{i,r}^{(n)}, i \in \mathbb{N}_n, r, n \in \mathbb{N}$, – момент надходження $(r + 1)$ – го об'єкту класу i у вищеписаній схемі пуассонізації. Тоді $Z_{i,r}^{(n)}$ є незалежною гамма-розподіленою випадковою величиною, $Z_{i,r}^{(n)} \sim \Gamma\left(r + 1, \frac{1}{n}\right)$. Для будь-якого фіксованого значення n послідовності $(Y_{i,r}^{(n)})$ та $(Z_{i,r}^{(n)})$ задані на одному просторі ймовірностей та пов'язані між собою наступним чином:

$$Z_{i,r}^{(n)} = \sum_{j=1}^{Y_{i,r}^{(n)}} E_j, \quad i \in \mathbb{N}_n, r \in \mathbb{N}. \quad (2.2)$$

Більше того, послідовності $(Y_{i,r}^{(n)}, i \in \mathbb{N}_n, r \in \mathbb{N})$ та $(E_j, j \in \mathbb{N})$ є незалежними, тоді з (2.1), минула послідовність визначається виключно значеннями M_k .

Позначимо

$$\psi_r^{(n)}(x) = \frac{x}{\frac{r}{n^{r+1}}}, \quad x \in \mathbb{R}. \quad (2.3')$$

Зокрема, при $r = 1$ маємо

$$\psi_1^{(n)}(x) = \frac{x}{\sqrt{n}}, x \in \mathbb{R}. \quad (2.3)$$

Тоді точковий процес нормованого $(r + 1)$ -го надходження об'єктів різних класів має наступний вигляд:

$$\xi_r^{(n)} = \sum_{i=1}^n \delta_{\psi_r^{(n)}(Y_{i,r}^{(n)})}. \quad (2.4)$$

У подальшому нам також знадобиться аналог $\xi_r^{(n)}$ в пуассонізованому випадку:

$$\eta_r^{(n)} = \sum_{i=1}^n \delta_{\psi_r^{(n)}(Z_{i,r}^{(n)})}. \quad (2.5)$$

2.3. Основний результат

Пригадаємо деякі основні означення, пов'язані зі збіжністю точкових процесів (див. [7], [8] або [10]).

Позначимо через $M_p(\mathbb{R})$ простір усіх локально фінітних точкових мір на $\mathbb{R}$. Будемо казати, що для $\mu, \mu_1, \mu_2, \dots \in M_p(\mathbb{R})$, μ_n збігається грубо до μ (позначається як $\mu_n \xrightarrow{v} \mu$), якщо $\int_{\mathbb{R}} f d\mu_n \rightarrow \int_{\mathbb{R}} f d\mu$ для кожної фінітної неперервної функції $f: \mathbb{R} \rightarrow [0, +\infty)$. Множина $M_p(\mathbb{R})$ з відповідною топологією може бути охарактеризована як повний сепарабельний метричний простір. Все вищесказане дозволяє розглядати відповідну збіжність за розподілом точкових процесів $\xi, \xi_1, \xi_2, \dots$, яка позначається $\xi_n \xrightarrow{vd} \xi$. Основним результатом в цій роботі, наведеним нижче в теоремі 2.1, є те, що точкові процеси $\xi_r^{(n)}$ збігаються в цьому сенсі до процесу Пуассона.

Теорема 2.1. *Нехай ξ_r є пуассонівським точковим процесом на $\mathbb{R}$ з мірою інтенсивності*

$$\lambda_r(dx) = \frac{x^r}{r!} \cdot \mathbb{I}\{x \geq 0\} dx.$$

Тоді $\xi_r^{(n)} \xrightarrow{vd} \xi_r$ при $n \rightarrow \infty$.

Розіб'ємо доведення теореми 2.1 на кілька простіших кроків. По-перше, ми доведемо подібний результат у пуассонізованому випадку.

Лема 2.1. $\eta_r^{(n)} \xrightarrow{vd} \xi_r$ при $n \rightarrow \infty$.

Доведення. Оскільки $Z_{i,r}^{(n)} \sim \Gamma\left(r + 1, \frac{1}{n}\right)$, то щільність $f_{i,r}^{(n)}$ має вигляд

$$f_{i,r}^{(n)}(x) = \frac{1}{n^{r+1}r!} x^r e^{-\frac{x}{n}} \mathbb{I}\{x \geq 0\}.$$

1) $r = 1$:

Тоді щільність $\tilde{f}_{i,1}^{(n)}$ з $\psi_1^{(n)}(Z_{i,1}^{(n)})$ набуде вигляду

$$\tilde{f}_{i,1}^{(n)}(x) = \sqrt{n} f_{i,1}^{(n)}(\sqrt{n}x) = \sqrt{n} \frac{1}{n^2} (\sqrt{n}x) e^{-\frac{(\sqrt{n}x)}{n}} \mathbb{I}\{x \geq 0\} = \frac{x}{n} e^{-\frac{x}{\sqrt{n}}} \mathbb{I}\{x \geq 0\}.$$

Звідси, для будь-якого $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} n \tilde{f}_{i,1}^{(n)}(x) = x \mathbb{I}\{x \geq 0\},$$

і, більше того, ця збіжність є рівномірною на обмежених множинах. Тоді, для кожної Борелевої множини B при $n \rightarrow \infty$ виконується наступне:

$$n \int_B \tilde{f}_{i,1}^{(n)}(x) dx \rightarrow \int_B x \mathbb{I}\{x \geq 0\} dx,$$

а, отже, (див. властивість 3.12 в [7]),

$$n \mathbb{P} \left\{ \psi_1^{(n)}(Z_{i,1}^{(n)}) \in \cdot \right\} \xrightarrow{v} \lambda_1(\cdot) \text{ на } \mathbb{R}.$$

2) r — довільне:

Розглянемо величини $\frac{Z_{i,r}^{(n)}}{n^\alpha}$ та покажемо, що $\alpha = \frac{r}{r+1}$:

Тоді $\tilde{f}_{i,r}^{(n)}(x) = n^\alpha f_{i,r}^{(n)}(n^\alpha x)$. Звідси

$$\begin{aligned}
\lim_{n \rightarrow \infty} n \tilde{f}_{i,r}^{(n)}(x) &= \frac{n \cdot n^\alpha}{n^{r+1} r!} (n^\alpha x)^{r+1} \cdot e^{-n^{\alpha-1} x} \cdot \mathbb{I}\{n^\alpha x \geq 0\} \\
&= \lim_{n \rightarrow \infty} \frac{n^{1+(r+1)(\alpha-1)}}{r!} \cdot e^{-n^{\alpha-1} x} \cdot \mathbb{I}\{x \geq 0\} = \langle 1 + (r+1)(\alpha-1) \\
&= 0 \Rightarrow \alpha = \frac{r}{r+1} < 1 \rangle = \frac{x^r}{r!} \cdot \mathbb{I}\{x \geq 0\}.
\end{aligned}$$

Варто зазначити, що при $r = 1$ отримуємо міркування, аналогічні до пункту 1).

Враховуючи незалежність $Z_{i,r}^{(n)}$ для різних i та добре відомий факт збіжності біноміального точкового процесу до пуассонівського, одержуємо доведення (див. властивість 3.21 в [7]). ■

На наступному кроці нам потрібно провести так звану процедуру депуассонізації, щоб перейти від збіжності $\eta_r^{(n)} \xrightarrow{vd} \xi_r$ до $\xi_r^{(n)} \xrightarrow{vd} \xi_r$. Такі методи депуассонізації зазвичай передбачають обмеження відстаней між випадковими елементами в пуассонізованому та депуассонізованому випадках (див. лему 1.4 в [11] щодо випадкових величин і, як її застосування, теорему 3.2 в [12] для точкових процесів). З того, що $\xi_r^{(n)}$ та $\eta_r^{(n)}$ пов'язані (2.2), ми можемо використати цю ідею для того, щоб оцінити, наскільки близькими є обидва ці процеси.

Лема 2.2. Для довільного проміжку $[a, b] \subset \mathbb{R}$ та для довільних $n \in \mathbb{N}$, $\varepsilon > 0$, $r > 1$,

$$\begin{aligned}
& \mathbb{P} \left\{ \xi_r^{(n)}[a, b] \neq \eta_r^{(n)}[a, b] \right\} \\
& \leq c_r \varepsilon^{-8} n^{\frac{8-3(r+1)}{r+1}} + \mathbb{P} \left\{ \eta_r^{(n)}[a - \varepsilon, a + \varepsilon] \geq 1 \right\} \\
& + \mathbb{P} \left\{ \eta_r^{(n)}[b - \varepsilon, b + \varepsilon] \geq 1 \right\},
\end{aligned} \tag{2.6}$$

де c_r — деяка стала.

Доведення. Ідея доведення (2.6) є доволі простою. Грубо кажучи, можливі два варіанти, коли $\xi_r^{(n)}[a, b] \neq \eta_r^{(n)}[a, b]$: або деякі точки $\xi_r^{(n)}$ значно відхилилися від відповідних точок $\eta_r^{(n)}$, або точки $\eta_r^{(n)}$ достатньо близькі до межі $[a, b]$. Перший доданок у правій частині (2.6) відповідає за першу причину, а решта за другу. Приступаємо до реалізації.

Зафіксуємо деяке $\varepsilon > 0$. Тоді, з (2.2) та (2.3'),

$$\begin{aligned}
& \mathbb{P} \left\{ \left| \psi_r^{(n)} \left(Z_{i,r}^{(n)} \right) - \psi_r^{(n)} \left(Y_{i,r}^{(n)} \right) \right| > \varepsilon \right\} = \mathbb{P} \left\{ \left| Z_{i,r}^{(n)} - Y_{i,r}^{(n)} \right| > \varepsilon n^{\frac{r}{r+1}} \right\} \\
& = \mathbb{P} \left\{ \left| \sum_{j=1}^{Y_{i,r}^{(n)}} (E_j - 1) \right| > \varepsilon n^{\frac{r}{r+1}} \right\}
\end{aligned}$$

для деякого $i \in \mathbb{N}_n$. Отже, за нерівністю Маркова останнє не перевищує

$$\Delta_{\varepsilon,n} = \left(\varepsilon n^{\frac{r}{r+1}} \right)^{-8} \mathbb{E} \left(\sum_{j=1}^{Y_{i,r}^{(n)}} (E_j - 1) \right)^8.$$

Оскільки $Y_{1,r}^{(n)} \sim \text{NegBin} \left(r + 1, \frac{1}{n} \right)$, а E_j — центровані н.о.р.в.в., то

$$\mathbb{E} \left(\sum_{j=1}^{Y_{i,r}^{(n)}} (E_j - 1) \right)^8 \leq c_r n^4,$$

тобто прямі обчислення можна обмежити зверху

$$\Delta_{\varepsilon,n} \leq c_r \varepsilon^{-8} n^{\frac{8-4(r+1)}{r+1}}, n \in \mathbb{N},$$

де c_r —деяка стала.

Тепер ми остаточно можемо обмежити ймовірність того, що деякі точки $\xi_r^{(n)}$ значно відхилилися від відповідних точок $\eta_r^{(n)}$: за властивістю субадитивності

$$\begin{aligned} \mathbb{P} \left\{ \left| \psi_r^{(n)} \left(Z_{i,r}^{(n)} \right) - \psi_r^{(n)} \left(Y_{i,r}^{(n)} \right) \right| > \varepsilon \text{ для деякого } i \in \mathbb{N}_n \right\} &\leq n \Delta_{\varepsilon,n} \\ &\leq c_r \varepsilon^{-8} n^{\frac{8-3(r+1)}{r+1}}. \end{aligned}$$

Зауважимо, що при $r > 1$: $\frac{8-3(r+1)}{r+1} < 0$.

Тоді (2.6) випливає з попередніх міркувань, що і доводить лему. ■

Для того, щоб розібратися з останніми двома доданками справа (2.6), нам знадобиться проста технічна лема про розподіл збіжних точкових процесів.

Лема 2.3. Нехай $X^{(n)}, n \in \mathbb{N}$, та X — точкові процеси на $\mathbb{R}$ такі, що $X^{(n)} \xrightarrow{vd} X$, та X має міру інтенсивності. Якщо $(I_n, n \in \mathbb{N})$ є спадною послідовністю інтервалів таких, що $I_n \downarrow I$ для деякого інтервалу I , тоді

$$\mathbb{P}\{X^{(n)}(I_n) \geq 1\} \rightarrow \mathbb{P}\{X(I) \geq 1\}.$$

Доведення. Доведення базується на перетворенні Скорохода (див. [8], с.41). Дивлячись на останнє, ми можемо припустити, що $(X_n, n \in \mathbb{N})$ та X задані на однаковому ймовірнісному просторі, тоді $X^{(n)} \xrightarrow{v} X$ м.н. Отже,

$$|X^{(n)}(I_n) - X(I)| \leq |X^{(n)}(I) - X(I)| + |X^{(n)}(I_n) - X^{(n)}(I)|. \quad (2.7)$$

З властивостей міри інтенсивності X випливає, що $X(\partial I) = 0$ м.н., і тоді перший доданок у правій частині (2.7) м.н. прямує до нуля при $n \rightarrow \infty$ через грубу збіжність. Щоб розібратися з другим доданком, зафіксуємо значення $N \in \mathbb{N}$. Для всіх $n \geq N$

$$|X^{(n)}(I_n) - X^{(n)}(I)| = X^{(n)}(I_n \setminus I) \leq X^{(n)}(I_N \setminus I),$$

тому права частина збігається м.н. до $X^{(n)}(I_N \setminus I)$. Іншими словами,

$$\lim_{n \rightarrow \infty} \sup |X^{(n)}(I_n) - X^{(n)}(I)| \leq X(I_N \setminus I) \text{ м.н.}$$

Спрямувавши $N \rightarrow \infty$, ми бачимо, що другий доданок правої частини (2.7) м.н. збігається до нуля, а отже, при $n \rightarrow \infty$ також.

Отже, $X^{(n)}(I_n) \rightarrow X(I)$ м.н., і тоді $X^{(n)}(I_n) \rightarrow X(I)$ за розподілом, що чітко доводить лему. ■

Нам залишилось провести фінальну частину доведення.

Доведення теореми 2.1. В (2.6) покладемо $\varepsilon = n^{-\frac{1}{25}}$. Тоді

$$\begin{aligned} & \mathbb{P} \left\{ \xi_r^{(n)}[a, b] \neq \eta_r^{(n)}[a, b] \right\} \\ & \leq c_r n^{\frac{8 - \frac{67}{25}(r+1)}{r+1}} + \mathbb{P} \left\{ \eta_r^{(n)} \left[a - n^{-\frac{1}{25}}, a + n^{-\frac{1}{25}} \right] \geq 1 \right\} \\ & \quad + \mathbb{P} \left\{ \eta_r^{(n)} \left[b - n^{-\frac{1}{25}}, b + n^{-\frac{1}{25}} \right] \geq 1 \right\}. \end{aligned}$$

Зауважимо, що $8 - \frac{67}{25}(r+1) < 0$ при $r > 1$.

З лем 2.1 та 2.3 випливає, що при $n \rightarrow \infty$

$$\mathbb{P} \left\{ \eta_r^{(n)} \left[a - n^{-\frac{1}{25}}, a + n^{-\frac{1}{25}} \right] \geq 1 \right\} \rightarrow \mathbb{P} \{ \xi_r \{a\} \geq 1 \} = 0,$$

і те ж саме виконується для b . Разом з вищеподаною нерівністю ми отримуємо наступне:

$$\mathbb{P} \left\{ \xi_r^{(n)}[a, b] \neq \eta_r^{(n)}[a, b] \right\} \rightarrow 0 \text{ при } n \rightarrow \infty. \quad (2.8)$$

Нехай $\mathcal{U}$ — об'єднання скінченних обмежених та замкнених відрізків на $\mathbb{R}$. Для кожного $U = [a_1, b_1] \cup \dots \cup [a_l, b_l] \in \mathcal{U}$ та $k \in \mathbb{N}$, з (2.8) випливає

$$\begin{aligned} \left| \mathbb{P} \left\{ \xi_r^{(n)}(U) = k \right\} - \mathbb{P} \left\{ \eta_r^{(n)}(U) = k \right\} \right| &\leq \mathbb{P} \left\{ \xi_r^{(n)}(U) \neq \eta_r^{(n)}(U) \right\} \\ &\leq \sum_{i=1}^l \mathbb{P} \left\{ \xi_r^{(n)}[a_i, b_i] \neq \eta_r^{(n)}[a_i, b_i] \right\} \rightarrow 0, \quad n \rightarrow \infty. \end{aligned}$$

Більше того, з леми 2.1

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \eta_r^{(n)}(U) = k \right\} = \mathbb{P} \{ \xi_r(U) = k \}.$$

Тоді остаточно маємо, що

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \xi_r^{(n)}(U) = k \right\} = \mathbb{P} \{ \xi_r(U) = k \}.$$

Зауважимо, що процес ξ_r простий, оскільки це процес Пуассона з дифузною мірою інтенсивності (див. власт. 6.9 в [5]). Отже, $\xi_r^{(n)} \xrightarrow{vd} \xi_r$ на $\mathbb{R}$ за теоремою 4.15 в [10]. ■

Розділ 3

Застосування основного результату

З основного результату магістерської дисертації випливає ряд важливих граничних результатів для величин $T_r^{(n)}$. Подамо їх у вигляді наступних теорем.

Теорема 3.1. *Справедливе наступне:*

$$n^{-\frac{r}{r+1}} T_r^{(n)} \xrightarrow{d} \rho, n \rightarrow \infty,$$

де

$$f_\rho(x) = \frac{x^r}{r!} e^{-\frac{x^{r+1}}{(r+1)!}} \cdot \mathbb{I}\{x \geq 0\}. \quad (3.1)$$

Доведення теореми 3.1. Оскільки $\xi_r^{(n)} \xrightarrow{dv} \xi_r$, то перша точка $\xi_r^{(n)}$ збігається за розподілом до першої точки ξ_r . Але перша точка $\xi_r^{(n)}$ – це $n^{-\frac{r}{r+1}} \min_{i=1, \dots, n} Y_{i,r}^{(n)} = n^{-\frac{r}{r+1}} T_r^{(n)}$. Розподіл першої точки ξ_r можна легко знайти, оскільки ξ_r є процесом Пуассона з мірою інтенсивності

$$\lambda_r(dx) = \frac{x^r}{r!} \cdot \mathbb{I}\{x \geq 0\} dx.$$

Покажемо, що величини $\frac{T_r^{(n)}}{n^{\frac{r}{r+1}}}$ мають узагальнений розподіл Релея:

$$\begin{aligned} F_{\xi_r}(x) &= \mathbb{P}\{\text{перша точка } \xi_r \leq x\} = 1 - \mathbb{P}\{\text{перша точка } \xi_r > x\} = \\ &= 1 - \mathbb{P}\{\text{кількість точок } \xi_r \text{ на } [0, x] \text{ рівна } 0\} \\ &= 1 - \mathbb{P}\{\xi_r[0, x] = 0\}. \end{aligned}$$

Оскільки $\xi_r[0, x] \sim \text{Pois} \left(\int_0^x \frac{t^r}{r!} dt \right) = \text{Pois} \left(\frac{x^{r+1}}{(r+1)!} \right)$, то $F_{\xi_r}(x) = \left(1 - e^{-\frac{x^{r+1}}{(r+1)!}} \right) \cdot \mathbb{I}\{x \geq 0\}$. Тоді щільність має вигляд

$$f_{\xi_r}(x) = \frac{x^r}{r!} e^{-\frac{x^{r+1}}{(r+1)!}} \cdot \mathbb{I}\{x \geq 0\}.$$

Перевіримо умову нормування:

$$\int_0^\infty f_{\xi_r}(x) dx = \int_0^\infty \frac{x^r}{r!} e^{-\frac{x^{r+1}}{(r+1)!}} dx = \int_0^\infty e^{-\frac{x^{r+1}}{(r+1)!}} d\left(\frac{x^{r+1}}{(r+1)!}\right) = 1.$$

Зауважимо, що у випадку $r = 1$ одержуємо щільність класичного розподілу Релея. ■

Наслідок 3.1. З теореми 3.1 випливає слабка збіжність величин $n^{-\frac{r}{r+1}} T_r^{(n)}$ до випадкової величини ρ , яка задана щільністю (3.1). З асимптотики за Ньюманом та Кламкіном, наведеної у вступі, випливає така збіжність:

$$\mathbb{E} n^{-\frac{r}{r+1}} T_r^{(n)} \underset{n \rightarrow \infty}{\sim} \frac{r+1}{\sqrt{(r+1)!}} \Gamma \left(1 + \frac{1}{r+1} \right).$$

Знайдемо математичне сподівання величини ρ та покажемо, що воно рівне правій частині вищеподаної збіжності.

$$\begin{aligned}
& \blacksquare \mathbb{E}\rho = \int_0^\infty \frac{x^{r+1}}{r!} e^{-\frac{x^{r+1}}{(r+1)!}} dx \\
& = \left| \begin{array}{l} u = x; \quad du = dx \\ dv = \frac{x^r}{r!} e^{-\frac{x^{r+1}}{(r+1)!}} dx; \\ v = \int \frac{x^r}{r!} e^{-\frac{x^{r+1}}{(r+1)!}} dx = \int e^{-\frac{x^{r+1}}{(r+1)!}} d\left(\frac{x^{r+1}}{(r+1)!}\right) = -e^{-\frac{x^{r+1}}{(r+1)!}} \end{array} \right| \\
& = \underbrace{-xe^{-\frac{x^{r+1}}{(r+1)!}} \Big|_0^\infty}_0 + \int_0^\infty e^{-\frac{x^{r+1}}{(r+1)!}} dx = \int_0^\infty e^{-\frac{x^{r+1}}{(r+1)!}} dx \\
& = \left| \begin{array}{l} \frac{x^{r+1}}{(r+1)!} = t \\ x = \sqrt[r+1]{(r+1)!} t^{1/(r+1)} \\ dx = \frac{1}{r+1} t^{1/(r+1)-1} dt \end{array} \right| \\
& = \int_0^\infty \sqrt[r+1]{(r+1)!} \frac{1}{r+1} t^{1/(r+1)-1} e^{-t} dt \\
& = \sqrt[r+1]{(r+1)!} \frac{1}{r+1} \Gamma\left(\frac{1}{r+1}\right).
\end{aligned}$$

Оскільки $z\Gamma(z) = \Gamma(z+1)$, то

$$\mathbb{E}\rho = \sqrt[r+1]{(r+1)!} \Gamma\left(1 + \frac{1}{r+1}\right),$$

що і треба було показати. $\blacksquare$

Розглянемо ще одне застосування теореми 2.1. Через величину $T_{r,k}^{(n)}$ позначимо перший момент часу, коли k класів міститимуть принаймні $(r+1)$ об'єктів. Наступна теорема описує розподіл таких величин при $n \rightarrow \infty$.

Теорема 3.2. *Виконується наступна збіжність:*

$$n^{-\frac{r}{r+1}} T_{r,k}^{(n)} \xrightarrow{d} \tau_{r,k}, n \rightarrow \infty,$$

де функція розподілу $\tau_{r,k}$ має вигляд

$$F_{\tau_{r,k}}(x) = 1 - \sum_{j=0}^{k-1} \frac{x^{(r+1)j}}{j! ((r+1)!)^j} e^{-\frac{x^{r+1}}{(r+1)!}}, x \geq 0.$$

Доведення теореми 3.2. За означенням

$$\begin{aligned} F_{n^{-\frac{r}{r+1}}T_{r,k}^{(n)}}(x) &= \mathbb{P}\left\{n^{-\frac{r}{r+1}}T_{r,k}^{(n)} \leq x\right\} \\ &= \mathbb{P}\left\{k - \text{та (за зростанням) величина } n^{-\frac{r}{r+1}}Y_{i,r}^{(n)} \leq x\right\} \\ &= \mathbb{P}\left\{k - \text{та точка } \xi_r^{(n)} \leq x\right\} \xrightarrow{d} \mathbb{P}\{k - \text{та точка } \xi_r \leq x\}, n \rightarrow \infty. \end{aligned}$$

Нехай

$$\begin{aligned} F_{\tau_{r,k}}(x) &= \mathbb{P}\{k - \text{та точка } \xi_r \leq x\} = 1 - \mathbb{P}\{k - \text{та точка } \xi_r > x\} = \\ &= 1 - \mathbb{P}\{\text{на } [0, x] \xi_r \text{ має } (k-1) \text{ точок, або } (k-2) \text{ точок, } \dots, \text{ або } 0 \text{ точок}\} \\ &= 1 - \sum_{j=0}^{k-1} \mathbb{P}\{\xi_r[0, x] = j\} = 1 - \sum_{j=0}^{k-1} \frac{e^{-\lambda_r[0, x]} (\lambda_r[0, x])^j}{j!} \\ &= 1 - \sum_{j=0}^{k-1} \frac{e^{-\frac{x^{r+1}}{(r+1)!}} \left(\frac{x^{r+1}}{(r+1)!}\right)^j}{j!} \\ &= 1 - \sum_{j=0}^{k-1} \frac{x^{(r+1)j}}{j! ((r+1)!)^j} e^{-\frac{x^{r+1}}{(r+1)!}}, x \geq 0. \end{aligned}$$

Знайдемо щільність $f_{\tau_{r,k}}(x), x \geq 0$. Для цього позначимо через $\lambda := \frac{x^{r+1}}{(r+1)!}$.

$$f_{\tau_{r,k}}(x) = \frac{dF_{\tau_{r,k}}(\lambda)}{d\lambda} \cdot \frac{d\lambda}{dx} = \frac{dF_{\tau_{r,k}}(x)}{d\lambda} \cdot \frac{x^r}{r!}.$$

Тепер обчислимо

$$\begin{aligned}
\frac{dF_{\tau_{r,k}}(\lambda)}{d\lambda} &= e^{-\lambda} + \sum_{j=0}^{k-1} \frac{e^{-\lambda} \lambda^j - e^{-\lambda} \lambda^{j-1}}{j!} \\
&= e^{-\lambda} + e^{-\lambda} \lambda - e^{-\lambda} + \frac{e^{-\lambda} \lambda^2}{2} - e^{-\lambda} \lambda + \dots + \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!} \\
&\quad - \frac{e^{-\lambda} \lambda^{k-2}}{(k-2)!} = \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!}.
\end{aligned}$$

Отже, повернувшись до заміни, одержуємо

$$f_{\tau_{r,k}}(x) = e^{-\frac{x^{r+1}}{(r+1)!}} \cdot \frac{x^{rk+k-1}}{r! (k-1)! ((r+1)!)^{k-1}}, x \geq 0.$$

Покажемо, що величина $\tau_{r,k}$ має узагальнений гамма-розподіл. Щільність такого розподілу має наступний вигляд:

$$f(x; a, d, p) = \frac{\left(\frac{p}{a^d}\right) x^{d-1} e^{-(x/a)^p}}{\Gamma(d/p)}, \quad a > 0, d > 0, p > 0.$$

У нашому випадку $d = k(r+1)$, $p = r+1$, $a = \sqrt[r+1]{(r+1)!}$. Підставимо дані значення у вищенаведену щільність та покажемо, що вона рівна $f_{\tau_{r,k}}$.

$$\begin{aligned}
&f\left(x; \sqrt[r+1]{(r+1)!}, k(r+1), r+1\right) \\
&= \frac{\left(r+1 / \left(\sqrt[r+1]{(r+1)!}\right)^{k(r+1)}\right) x^{k(r+1)-1} e^{-\left(x / \sqrt[r+1]{(r+1)!}\right)^{r+1}}}{\Gamma(k(r+1) / r+1)} \\
&= \frac{\frac{r+1}{((r+1)!)^k} x^{k(r+1)-1}}{\Gamma(k)} \cdot e^{-\frac{x^{r+1}}{(r+1)!}} = \frac{x^{rk+k-1} \cdot e^{-\frac{x^{r+1}}{(r+1)!}}}{r! (k-1)! ((r+1)!)^{k-1}} \\
&= f_{\tau_{r,k}}(x),
\end{aligned}$$

що і потрібно було довести. ■

Для доведення ще одного результату сформулюємо одну допоміжну теорему (див. теорему 7.4 в [5]).

Теорема 3.3. Припустимо, що ξ процес Пуассона на $\mathbb{R}_+ \times \mathbb{Y}$ з σ -фінітною мірою інтенсивності λ такою, що $\mathbb{P}(\xi(\cdot \times \mathbb{Y}) \in N^+) = 1$. Для будь-якого $n \in \mathbb{N}$, покладемо $T_n := T_n(\eta)$, де $\eta := \xi(\cdot \times \mathbb{Y})$. Розглянемо марковану послідовність Y_n . Тоді для будь-якої $f \in \mathbb{R}_+((\mathbb{R}_+ \times Y)^n)$, $n \in \mathbb{N}$, виконується наступне:

$$\begin{aligned} & \mathbb{E}[\mathbb{I}\{T_n < \infty\} g(T_1, Y_1, \dots, T_n, Y_n)] \\ &= \int \mathbb{I}\{0 < t_1 < \dots < t_n\} \times g(t_1, y_1, \dots, t_n, y_n) \exp[-\lambda((0, t_n] \\ & \times Y)] \lambda^n(d(t_1, y_1, \dots, t_n, y_n)). \end{aligned}$$

Теорема 3.4. Для довільного $r \geq 1$ справедлива така збіжність:

$$\frac{T_{r,2}^{(n)} - T_{r,1}^{(n)}}{n^{\frac{r}{r+1}}} \xrightarrow{d} \theta,$$

де щільність величини θ має вигляд

$$f_\theta(x) = \int_0^\infty \frac{t_1^r}{r!} e^{-\frac{(x+t_1)^{r+1}}{(r+1)!}} \frac{(x+t_1)^r}{r!} dt_1.$$

Доведення теореми 3.4. Аналогічно до попередніх теорем

$$\frac{T_{r,2}^{(n)} - T_{r,1}^{(n)}}{n^{\frac{r}{r+1}}} = \frac{Y_{r,(2)}^{(n)} - Y_{r,(1)}^{(n)}}{n^{\frac{r}{r+1}}},$$

де $n^{-\frac{r}{r+1}} Y_{r,(i)}^{(n)} - i$ — та точка $\xi_r^{(n)}$, $i = 1, 2$. Тоді за теоремою 2.1 різниця між другою та першою точками $\xi_r^{(n)}$ прямує за розподілом до різниці між

другою та першою точками процесу ξ_r при $n \rightarrow \infty$. Позначимо цей граничний розподіл через θ та знайдемо його.

Скориставшись позначеннями теореми 3.3, у нашому випадку матимемо $Y = \{1\}$, $g(T_1, Y_1, \dots, T_n, Y_n) = g(T_2 - T_1)$, $\lambda((0, t_n]) \times Y = \frac{t_2^{r+1}}{(r+1)!}$.

Тоді, з одного боку, для довільної функції g

$$\mathbb{E}g(T_2 - T_1) = \iint_{0 < t_1 < t_2} g(t_2 - t_1) e^{-\frac{t_2^{r+1}}{(r+1)!}} \frac{t_1^r}{r!} dt_1 \frac{t_2^r}{r!} dt_2,$$

а з іншого,

$$\mathbb{E}g(T_2 - T_1) = \int_0^\infty g(x) f_{T_2 - T_1}(x) dx,$$

де $f_{T_2 - T_1}$ — щільність розподілу $T_2 - T_1$.

Тоді з першої рівності

$$\begin{aligned} \mathbb{E}g(T_2 - T_1) &= \iint_{0 < t_1 < t_2} g(t_2 - t_1) e^{-\frac{t_2^{r+1}}{(r+1)!}} \frac{t_1^r}{r!} dt_1 \frac{t_2^r}{r!} dt_2 \\ &= \int_0^\infty \frac{t_1^r}{r!} dt_1 \int_{t_1}^\infty g(t_2 - t_1) e^{-\frac{t_2^{r+1}}{(r+1)!}} \frac{t_2^r}{r!} dt_2 = \langle t_2 - t_1 = x \rangle_{t_2 = x + t_1} = \\ &= \int_0^\infty \frac{t_1^r}{r!} dt_1 \int_0^\infty g(x) e^{-\frac{(x+t_1)^{r+1}}{(r+1)!}} \frac{(x+t_1)^r}{r!} dx \\ &= \int_0^\infty g(x) \left[\int_0^\infty \frac{t_1^r}{r!} e^{-\frac{(x+t_1)^{r+1}}{(r+1)!}} \frac{(x+t_1)^r}{r!} dt_1 \right] dx. \end{aligned}$$

Таким чином, з другої рівності випливає, що

$$f_{\theta}(x) = \int_0^{\infty} \frac{t_1^r}{r!} e^{-\frac{(x+t_1)^{r+1}}{(r+1)!}} \frac{(x+t_1)^r}{r!} dt_1. \blacksquare$$

З останньої теореми випливає один важливий наслідок, пов'язаний з математичним сподіванням різниці між нормованими першими моментами надходжень принаймні $(r + 1)$ об'єктів, які належать двом різним довільним класам. Цю величину можна характеризувати як середню тривалість проміжку між першим моментом появи $(r + 1)$ -го представника якогось класу та наступним таким моментом.

Отже,

$$\begin{aligned} \mathbb{E}(T_2 - T_1) &= \iint_{0 < t_1 < t_2} (t_2 - t_1) e^{-\frac{t_2^{r+1}}{(r+1)!}} \frac{t_1^r}{r!} dt_1 \frac{t_2^r}{r!} dt_2 = \\ &= \int_0^{\infty} \frac{t_1^r}{r!} dt_1 \int_{t_1}^{\infty} (t_2 - t_1) e^{-\frac{t_2^{r+1}}{(r+1)!}} \frac{t_2^r}{r!} dt_2. \end{aligned} \quad (3.2)$$

Нескладними математичними обчисленнями можна показати, що значення шуканого інтеграла рівне

$$\begin{aligned} \mathbb{E}(T_2 - T_1) &= \mathbb{E}T_2 - \mathbb{E}T_1 = \\ &= \sqrt[r+1]{(r+1)!} \Gamma\left(2 + \frac{1}{r+1}\right) - \sqrt[r+1]{(r+1)!} \Gamma\left(1 + \frac{1}{r+1}\right) \\ &= \frac{\sqrt[r+1]{(r+1)!}}{r+1} \Gamma\left(\frac{r+2}{r+1}\right). \end{aligned}$$

Наведемо кілька значень інтеграла (3.2) у випадку різних r та покажемо, що зі зростанням n дані математичні сподівання прямують до нескінченності.

```

> r<-3
>
> (((factorial(r))^(1/(r+1)))/(r+1))*gamma((r+2)/(r+1))
[1] 0.3546491

> r<-5
>
> (((factorial(r+1))^(1/(r+1)))/(r+1))*gamma((r+2)/(r+1))
[1] 0.4629003

> r<-8
>
> (((factorial(r+1))^(1/(r+1)))/(r+1))*gamma((r+2)/(r+1))
[1] 0.4363581

> r<-13
>
> (((factorial(r+1))^(1/(r+1)))/(r+1))*gamma((r+2)/(r+1))
[1] 0.4160886

```

Наприклад, у випадку $r = 13$ маємо

$$\mathbb{E}\left(T_{r,2}^{(n)} - T_{r,1}^{(n)}\right) = 0.4160886 \cdot n^{\frac{13}{14}},$$

що є достатньо великим числом при великих n .

Розділ 4

Візуалізація даних в середовищі RStudio

Здійснимо комп'ютерну симуляцію узагальненої задачі про дні народження за допомогою імітаційного моделювання в середовищі RStudio.

4.1. Візуалізація I

Згенеруємо вибірку об'ємом 2000 елементів, об'єктами якої будуть перші моменти часу, коли деякий з класів з'явився у $(r + 1)$ – ий раз. Для реалізації зафіксуємо значення r та n . Величина n описує загальну кількість класів у нашій задачі.

- 1) Нехай $r = 1, n = 1000$. Код відповідної програми буде мати наступний вигляд:

```
options(max.print=1000000)
> r<-2
> N<-2000
> steps<-vector()
> for (i in 1:N) {
+   arr<-rep(0, 1000)
+   step<-0
+   repeat {
+     randn<-sample(1:1000,1,replace=TRUE)
+     arr[randn]<-arr[randn]+1
+     step<-step+1
+     if (arr[randn] == r){
+       break
+     }
+   }
+   steps[i]<-step
+ }
>
> steps
[1] 22 26 44 31 18 54 31 26 65 68 49 31
29 64 30 42 71 17 81 44 44 46
[23] 34 16 31 19 24 52 24 53 29 37 44 58
31 12 49 93 24 52 17 25 26 14
...
[1959] 41 33 35 40 36 25 35 26 36 53 25 55
9 20 56 30 50 22 48 38 59 53
```

[1981] 82 15 25 9 63 49 36 60 49 54 68 46
114 29 32 23 64 54 47 48

(Для стислості викладу всю вибірку наводити не будемо.)

Для наочності одержаних результатів побудуємо гістограму частот. Для цього виконаємо наступне нормування для значень нашої вибірки та відсортуємо її:

```
> s1 <- steps / (1000^((r-1)/r))
> s <- sort(s1)
> # Гістограма частот
f1 <- function(t) {t*exp(-(t^2)/2)} # щільність розподілу
Релея
> plot(f1)
> # Гістограма частот
> hist(s1, freq = F, col = 'green4', border = 'black',
main = "Гістограма частот")
> curve(f1, add = T)
```

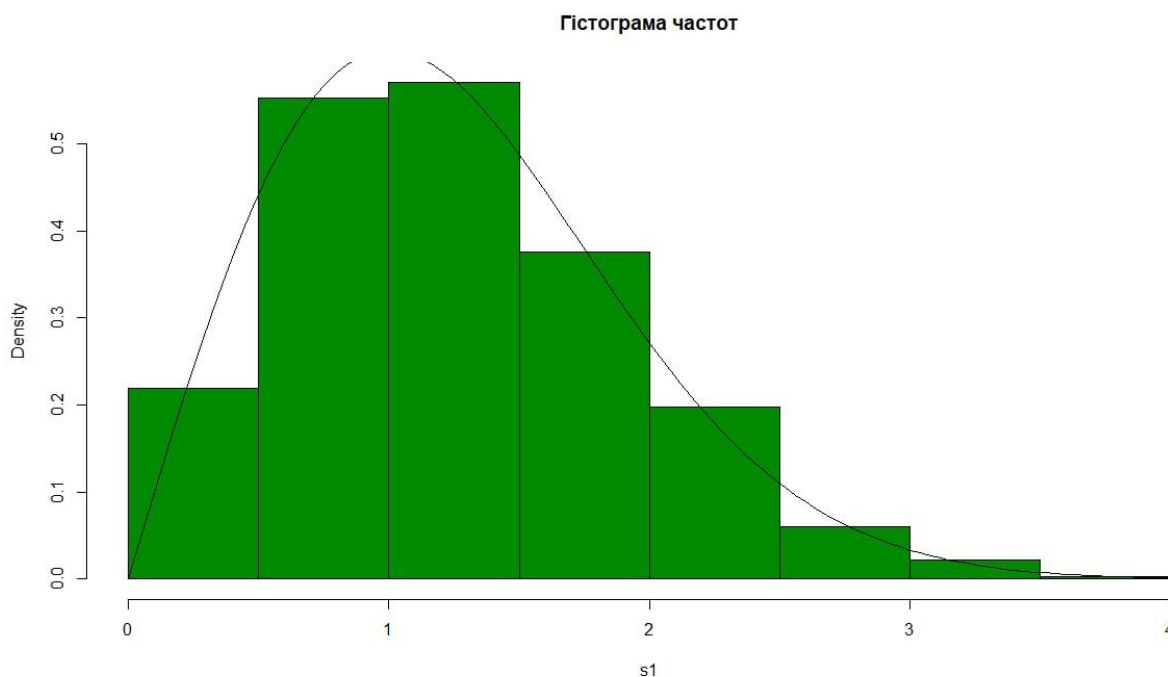


Рис. 4.1. Гістограма частот у випадку $r = 1, n = 1000$

Перевіримо узгодженість даної вибірки з теоретичним розподілом, а саме з розподілом Релея з параметром $\sigma = 1$ за допомогою критерія Хі-квадрат Пірсона.

```
> #Тест Хі-квадрат Пірсона
```

Обчислимо кількість інтервалів розбиття нашої вибірки за допомогою формули Стерджеса:

```
> #кількість інтервалів
> k <- round(log(N, base = 2) + 1)
> k
```

```
[1] 12
```

Довжини інтервалів відповідно будуть рівними

```
> #довжини інтервалів
> d<-(s[2000]-s[1])/11
> d
```

```
[1] 0.3852229
```

Утворимо послідовність, що складається з відповідних інтервалів розбиття:

```
> min1 <- min(s)
> max1 <- max(s)
> seq1 <- seq(min1, max1, length.out = k)
```

Підрахуємо частоти потрапляння в кожний з інтервалів, а також теоретичні ймовірності потрапити у відповідний інтервал. Для цього будемо користуватися наступною формулою:

$$\mathbb{P}\{\alpha < x < \beta\} = F(\beta) - F(\alpha),$$

де $F(x) = 1 - e^{-\frac{x^2}{2}}, x \geq 0$, — функція розподілу Релея.

Тоді одержимо наступні частоти та ймовірності:

```
> a <- 0
> c <- 0
> p <- 0
> #частоти потрапляння в інтервал та теоретичні
ймовірності потрапляння
> for (a in 1:(k-1)) {
+   c[a] <- length(s[s >= seq1[a] & s <= seq1[a+1]])
+   p[a] <- exp(-((seq1[a])^r)/factorial(r))-exp(-
((seq1[a+1])^r)/factorial(r))
+ }
> sum(c) #перевіряємо, чи всі елементи враховано
[1] 2000
> c #частоти
[1] 170 334 417 389 295 200 118 46 20 8 3
> p #ймовірності
```

```
[1] 0.081215047 0.173589402 0.211531477 0.196027042
0.148654877 0.095058658 0.052044500
[8] 0.024612550 0.010109957 0.003620515 0.001133329
```

Обчислимо $N \cdot p_i, i = \overline{1,11}$:

```
>n_p <- N*p
>n_p
[1] 162.430093 347.178804 423.062954 392.054083
297.309755 190.117316 104.089000 49.225099
[9] 20.219914 7.241030 2.266658
```

Оскільки повинна виконуватись умова $N \cdot p_i > 5$, об'єднаємо останні два інтервали в один:

```
> p1<-c(n_p[1]:n_p[9], n_p[10]+n_p[11])
> c1<-c(c[1]:c[9], c[10]+c[11])
```

Статистика критерія Пірсона має вигляд

$$\chi^2 = \sum_{i=1}^n \frac{(n_i - np_i)^2}{np_i},$$

де n_i — частоти потрапляння в кожен з інтервалів, p_i — теоретичні ймовірності потрапляння у відповідний інтервал, n — загальна кількість розбиттів, $i = \overline{1, n}$.

Підрахуємо її значення у нашому випадку:

```
> chi<-sum(((c1-p1)^2)/(p1))
> chi
[1] 3.80246
```

Критичне значення критерія Пірсона для $n = 10$ інтервалів розбиття, $q = 1$ ступенів свободи, при рівні значущості $\alpha = 0.05$ рівне $\chi_{n-q-1,0.05}^2 = \chi_{8,0.05}^2 = 15.507$. Оскільки значення статистики у нашому випадку дорівнює $3.80246 = \chi^2 < \chi_{8,0.05}^2 = 15.507$, то нема підстав відхилити гіпотезу про узгодженість розподілу одержаної вибірки з розподілом Релея.

Побудуємо емпіричну функцію для отриманої вибірки:

```

> f2<-function(t) {1-(exp(-(t^2)/2))} #функція розподілу
Релея
> plot(f2)
> #Емпірична функція розподілу
> plot(ecdf(s), pch=".", xlim=range(s), main="Емпірична
функція розподілу", xlab="", ylab="")
> curve(f2,add = T, col = 'red1', main = "lwd=20")

```

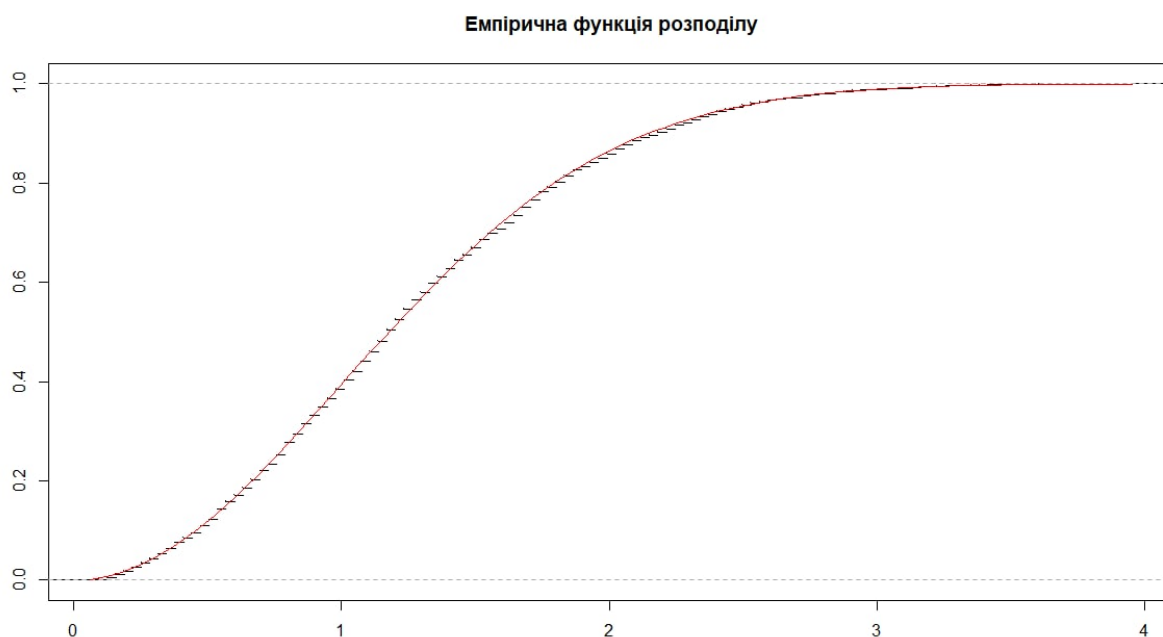


Рис. 4.2. Емпірична та теоретична функції розподілу у випадку $r = 1, n = 1000$

- 2) Нехай $r = 4$. Збільшимо кількість класів до $n = 30000$. В такому випадку одержимо наступну вибірку:

```

options(max.print=1000000)
> r<-5
> N<-2000
> steps1<-vector()
> for (i in 1:N) {
+   arr<-rep(0, 30000)
+   step<-0
+   repeat {
+     randn<-sample(1:30000,1,replace=TRUE)
+     arr[randn]<-arr[randn]+1
+     step<-step+1
+     if (arr[randn] == r){
+       break
+     }
+   }
+   steps1[i]<-step
+ }
>
> steps1

```

```

[1] 11506 5230 9581 9405 12791 8738 15001 10851
4532 9873 10408 12255 12281 9813
[15] 4888 11031 13638 11401 8233 12763 8406
9907 8079 7139 11253 12855 9402 9160
[29] 7271 12046 8875 9390 9030 12009 11815
12619 4013 12283 11617 10688 4518 8893
[43] 8587 9796 8641 12069 11642 12919 6620
8311 13895 10371 9778 11626 10078 8464
[57] 7195 9922 11751 9165 8870 10046 9345
9042 10851 13361 8711 11801 13089 10677
[71] 11298 11397 14940 10155 12801 11037 11985
8237 9704 9397 8054 8867 11332 9285

```

...

Відсортуємо її з урахуванням потрібного нормування:

```

s2<-steps/(30000^((r-1)/r))
> s<-sort(s2)

```

Емпірична функція для даної вибірки має наступний вигляд:

```

f2<-function(t) {1-(exp(-(t^r)/factorial(r)))}
#функція узагальненого розподілу Релея
> plot(f2)
> #Емпірична функція розподілу
> plot(ecdf(s), pch=".", xlim=range(s),
main="Емпірична функція розподілу", xlab="", ylab="")
> curve(f2,add = T, col = 'red1', main = "lwd=20")

```

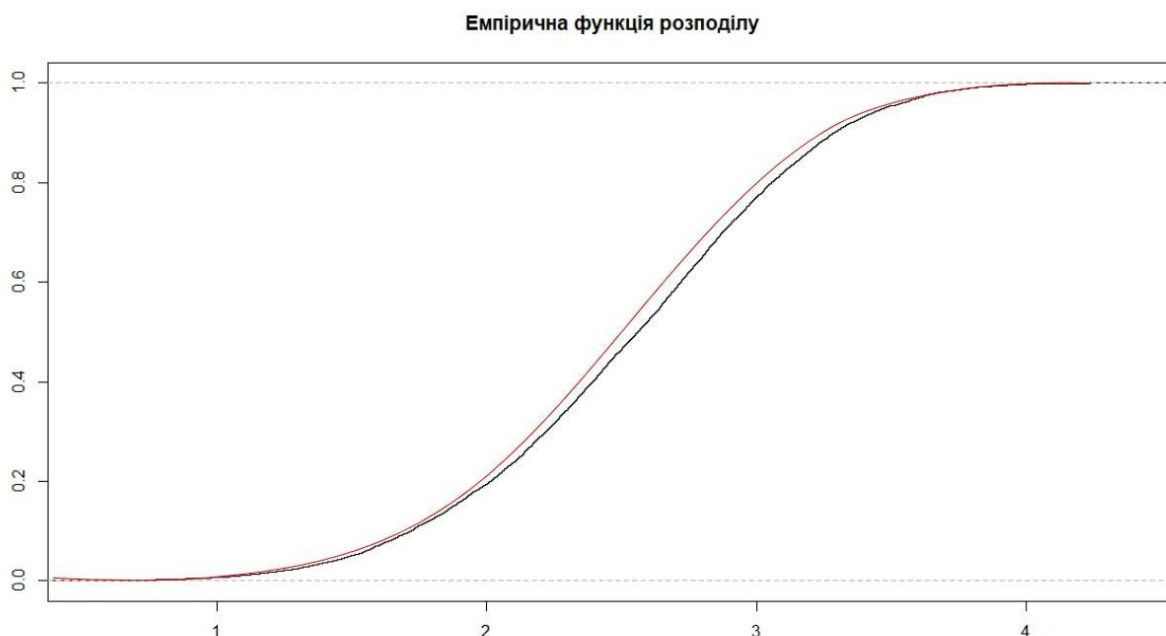


Рис. 4.3. Емпірична та теоретична функції розподілу у випадку $r = 4, n = 30000$

Побудуємо гістограму частот:

```
f1<-function(t)      {((t^(r-1))/factorial(r-1))*exp(-
(t^r)/factorial(r))} #щільність узагальненого розподілу Релея
> plot(f1)
> #Гістограма частот
> hist(s1, freq = F, col = 'green4', border = 'black',
main = "Гістограма частот")
> curve(f1,add = T)
> hist(steps1, freq = T, col = 'green4', border = 'black',
main = "Гістограма частот",breaks = 20)
```

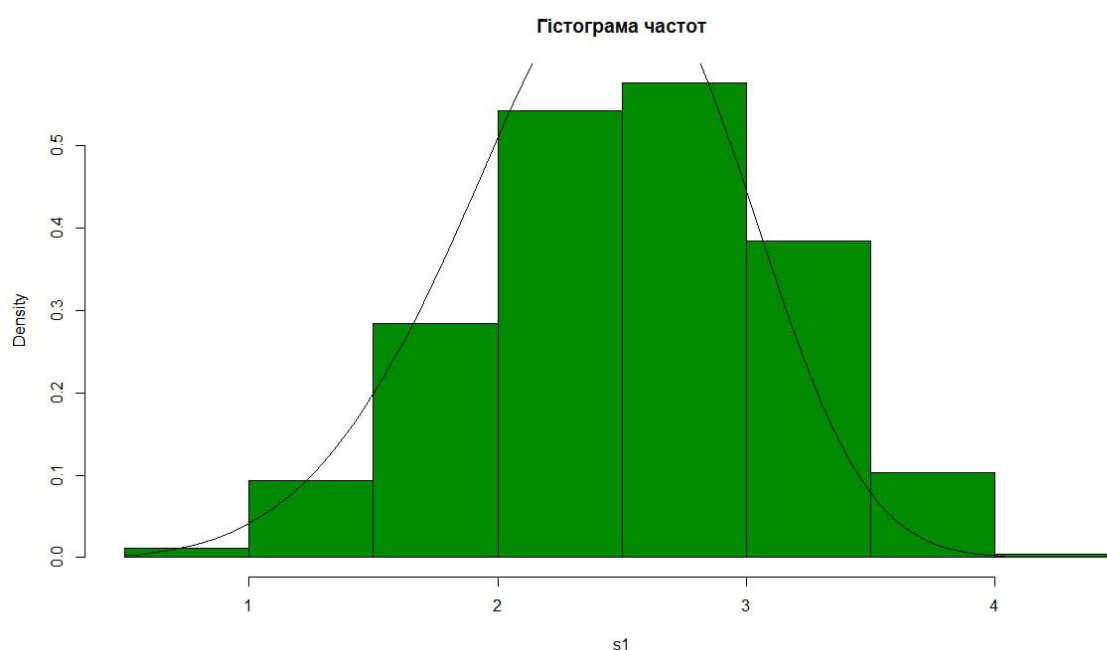


Рис. 4.4. Гістограма частот у випадку $r = 4, n = 30000$

3) Нехай $r = 19, n = 70000$. Тоді одержуємо наступну гістограму:

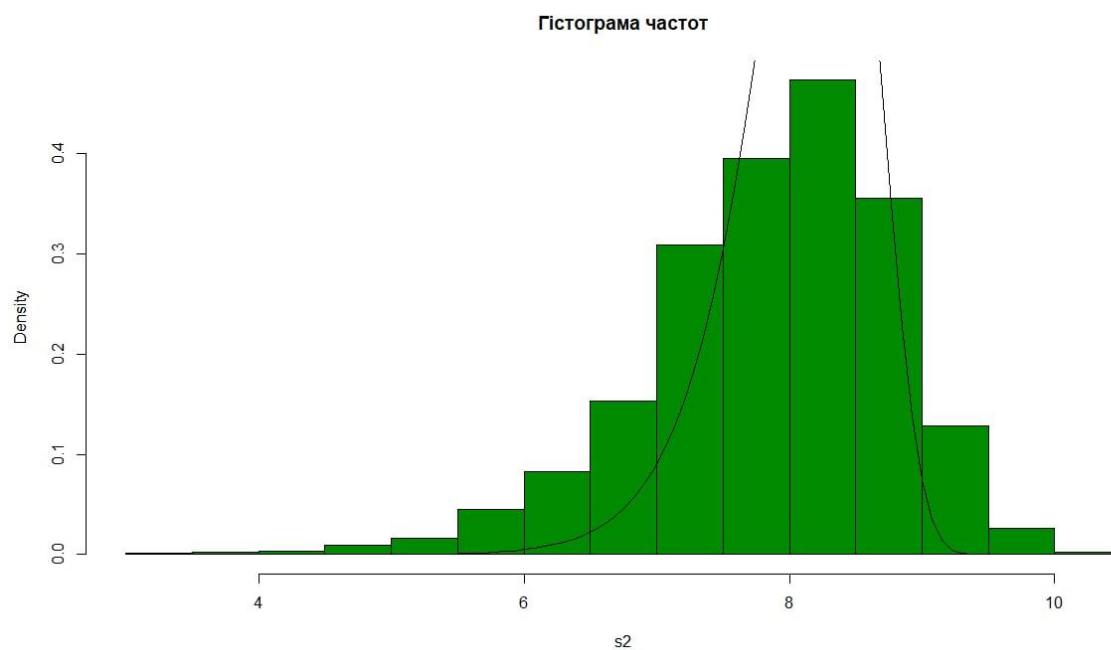


Рис. 4.5. Гістограма у випадку $r=19$, $n=70000$

4.2. Візуалізація II

Згенеруємо вибірку об'ємом 2000 елементів, об'єктами якої будуть перші моменти часу, коли k класів з'явилися принаймні $(r + 1)$ – ий разів, тобто нормовані величини $T_{r,k}^{(n)}$. Аналогічно до попереднього, для реалізації зафіксуємо значення r та n .

1) $r = 1, k = 5, n = 15000$:

```
options(max.print=1000000)
> r<-1
> N<-2000
> k<-5
> steps<-vector()
> for (i in 1:N) {
+   arr<-rep(0, 15000)
+   step<-0
+   box<-0
+   repeat {
+     randn<-sample(1:15000,1,replace=TRUE)
+     arr[randn]<-arr[randn]+1
+     step<-step+1
+     if (arr[randn] == r+1) {
+       box<-box+1
+     }
+     if (box == k){
+       break
+     }
+   }
+   steps[i]<-step
+ }
> steps
[1] 127 83 152 67 84 118 132 82 105 124 109 83 106
83 74 87 127 76 65 117 106 99
[23] 66 115 120 93 134 134 117 110 123 82 85 115 85
108 96 83 110 87 73 89 112 115
...
> s1<-steps/(15000^(r/(r+1)))
> s<-sort(s1)
> f<-function(t) {((exp(-(t^2)/2))/384)*(t^9)} #щільність
> #Гістограма частот
> hist(s, freq = F, col = 'blue1', border = 'black', main
= "Гістограма частот")
> curve(f,add = T)
```

Тоді гістограма має наступний вигляд:

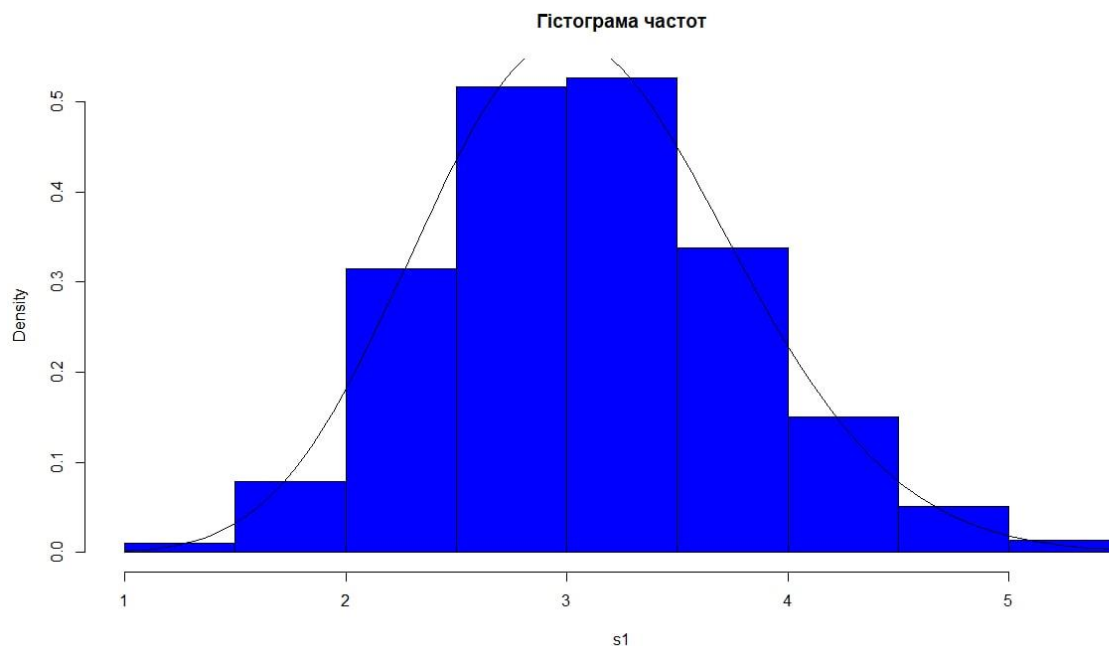


Рис. 4.6. Гістограма частот у випадку $r = 1, k = 5, n = 15000$

Перевіримо дану вибірку на узгодженість з теоретичним законом розподілу за критерієм Пірсона при рівні значущості $\alpha = 0.05$.

```
> l<-round(log(N, base = 2) + 1)
> min1 <- min(s)
> max1 <- max(s)
> seq1 <- seq(min1, max1, length.out = l)
> a <- 0
> c <- 0
> v <- 0
>
>
> #частоти потрапляння в інтервал та теоретичні
ймовірності потрапляння
> for (a in 1:(l-1)) {
+   c[a] <- length(s[s >= seq1[a] & s <= seq1[a+1]])
+   v[a] <- -F(seq1[a])+F(seq1[a+1])
+ }
> sum(c)
[1] 2000
> c
[1] 15 64 189 355 434 380 291 143 90 27 12
> v
[1] 0.007763429 0.036093391 0.097204250 0.171775107
0.214630342 0.198964805 0.141395429
[8] 0.078843450 0.035090235 0.012627099 0.003710682
> N*v
[1] 15.526858 72.186783 194.408500 343.550213
429.260683 397.929611 282.790858 157.686900
[9] 70.180469 25.254199 7.421364
```

```
> chi<-sum(((c-N*v)^2)/(N*v))
> chi
[1] 12.48752
```

Критичне значення статистики критерія становить $18.307 > 12.48752$, тобто у нас нема підстав на те, щоб відхилити гіпотезу про узгодженість теоретичного та емпіричного законів розподілу.

2) $r = 4, k = 25, n = 1000$:

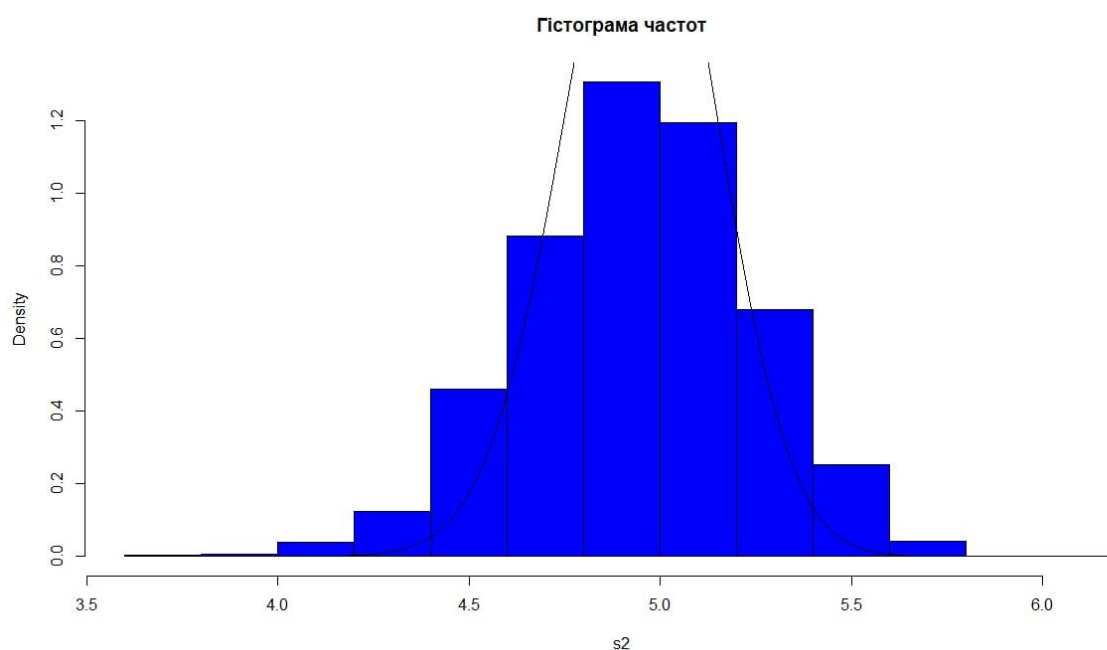


Рис. 4.7. Гістограма частот у випадку $r = 4, k = 25, n = 1000$

3) $r = 8, k = 20, n = 1000$:

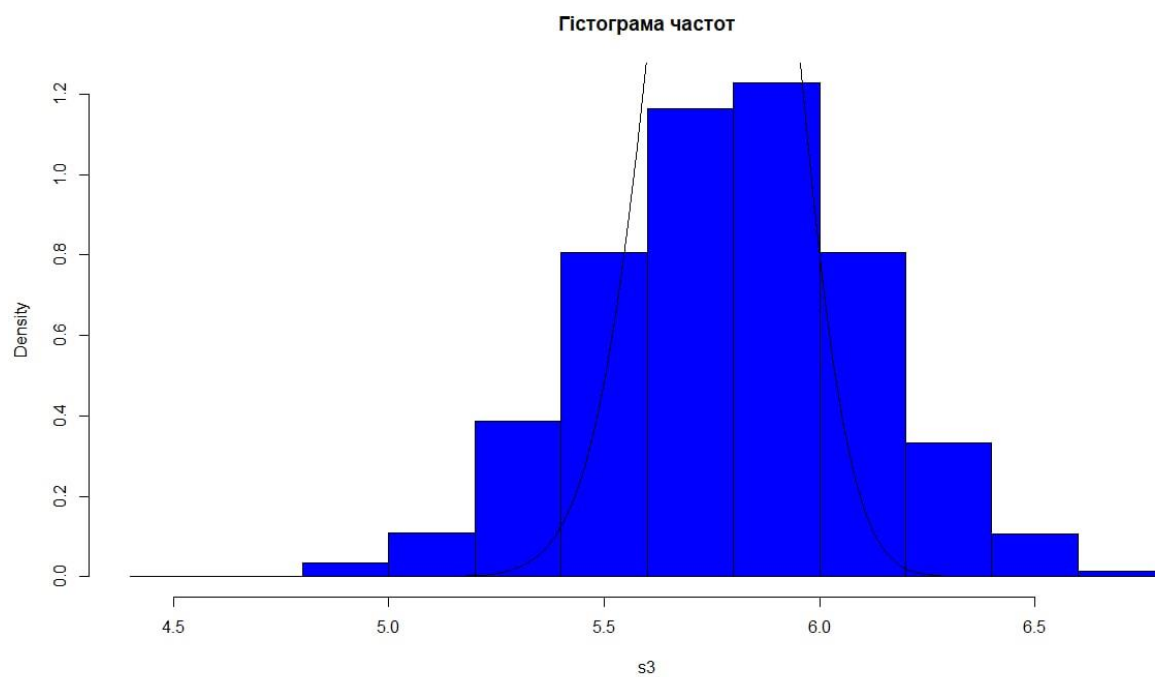


Рис. 4.8. Гістограма частот у випадку $r = 8, k = 20, n = 1000$

4.3. Візуалізація III

Згенеруємо вибірку об'ємом 2000 елементів, об'єктами якої будуть різниці між нормованими першими моментами надходжень принаймні $(r + 1)$ об'єктів, які належать двом різним довільним класам.

1) $r = 1, n = 1000$:

```
> options(max.print=1000000)
> r<-1
> N<-2000
> k_1<-1
> k_2<-2
> steps_1<-vector()
> steps_2<-vector()
> numb<-vector()
> for (i in 1:N) {
+   arr<-rep(0, 1000)
+   step<-0
+   box<-0
+   repeat {
+     randn<-sample(1:1000,1,replace=TRUE)
+     arr[randn]<-arr[randn]+1
+     step<-step+1
+     numb[step]<-randn
+     if (arr[randn] == r+1) {
+       box<-box+1
+       if (box == k_1){
+         step_1<-step
+       }
+     }
+     if (box == k_2){
+       break
+     }
+   }
+   steps_1[i]<-step_1
+   steps_2[i]<-step
+ }
>
> steps_1
[1] 39 23 57 42 17 76 122 16 51 62 35 29 39
57 30 57 76 23 58 38 58 51
[23] 53 46 37 23 34 38 55 51 73 82 57 20 31
15 32 62 15 76 26 26 23 16
f <- function(x) (sqrt(pi/2)*(erf(Inf)-erf(x/sqrt(2))))
#щільність
> # integral
> hist(diff,freq = F, col='orange2', main="Гістограма
частот")
> x <- seq(0,max_ch,.001)
> #Гістограма частот
> curve(f,add=T)
```

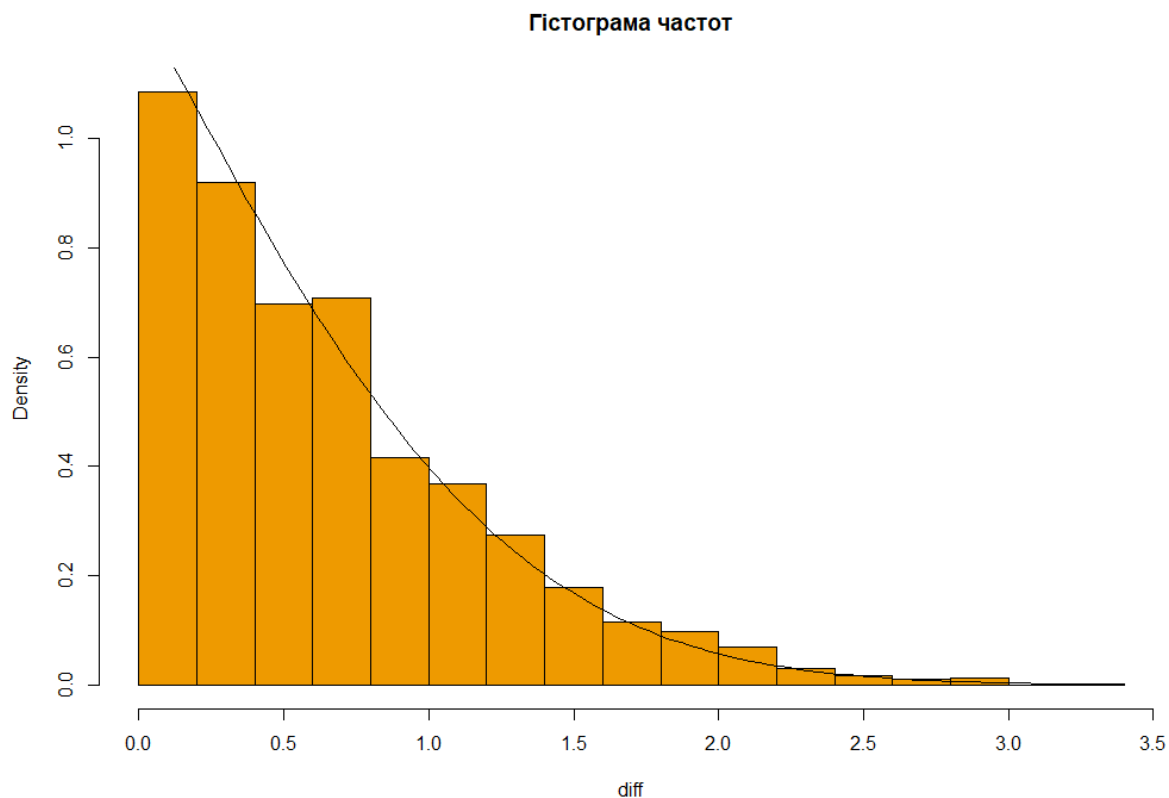


Рис. 4.9. Гістограма частот у випадку $r = 1, n = 1000$

Перейдемо до перевірки гіпотези про рівність теоретичного та емпіричного законів розподілу за критерієм Хі-квадрат.

```
> max_ch<-max(diff)
>
> f <- function(x) (sqrt(pi/2)*(erf(Inf)-erf(x/sqrt(2))))
# order of arguments reversed
>
> hist(diff,freq = F, col='orange2', main="Гістограма
частот")
> x <- seq(0,max_ch,.001)
> #Гістограма частот
> curve(f,add=T)
>
> Fr<-function(t) {integrate(f,0,t)$value}
> l<-round(log(N, base = 2) + 1)
> min1 <- min(s)
> max1 <- max(s)
> seq1 <- seq(min1, max1, length.out = l)
> a <- 0
> c <- 0
> v <- 0
```



```

> #частоти потрапляння в інтервал та теоретичні
ймовірності потрапляння
> for (a in 1:(l-1)) {
+   c[a] <- length(s[s >= seq1[a] & s <= seq1[a+1]])
+   v[a] <- -Fr(seq1[a])+Fr(seq1[a+1])
+ }
> sum(c)
[1] 2000
> c
[1] 631 429 334 229 161 96 52 38 16 10 4
> v
[1] 0.282912966 0.219059828 0.161708694 0.113497190
0.075565002 0.047632288
[7] 0.028380736 0.015962356 0.008464970 0.004228546
0.001988111
> N*v
[1] 565.825931 438.119655 323.417388 226.994380
151.130004 95.264577
[7] 56.761473 31.924711 16.929940 8.457093 3.976223
> chi<-sum(((c-N*v)^2)/(N*v))
> chi
[1] 10.59936

```

Критичне значення статистики – 18.307. Оскільки одержане значення 10.59936 не перевищує критичного, то при рівні значущості $\alpha = 0.05$ у нас нема жодних підстав на те, щоб відхилити нульову гіпотезу про однаковий розподіл.

2) $r = 5, n = 10000$:

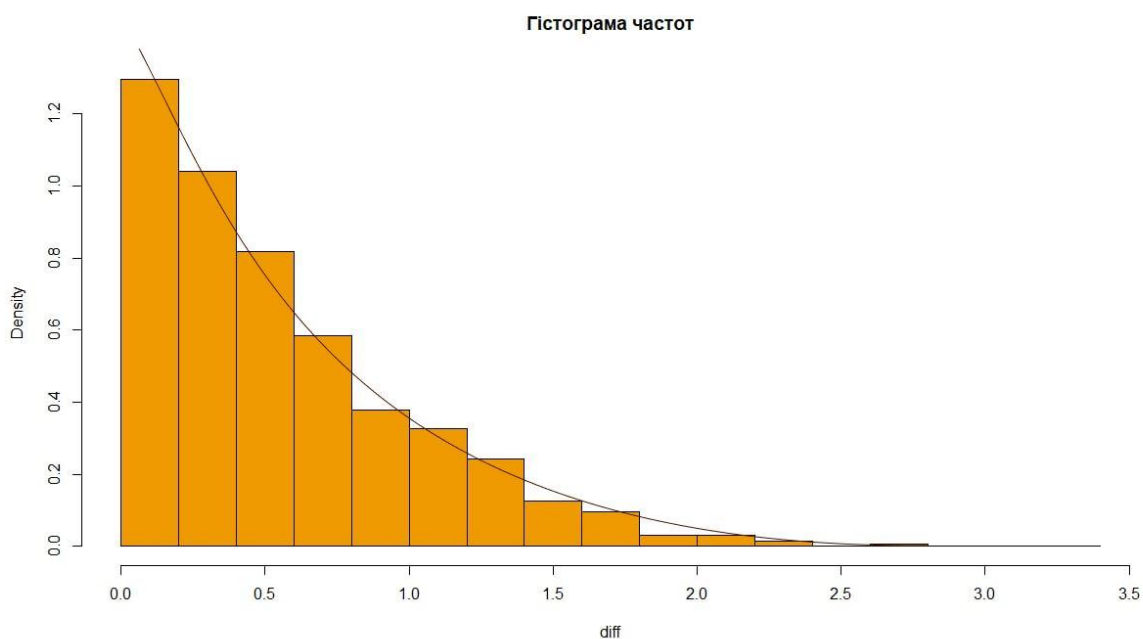


Рис. 4.10. Гістограма частот у випадку $r = 5, n = 10000$

Висновки

У даній магістерській дисертації було наведено узагальнення однієї з класичних задач теорії ймовірностей – задачі про дні народження. Було розглянуто послідовність об'єктів довільної природи, що один за одним надходять до спостерігача і кожен з яких незалежно від інших з рівними ймовірностями належить до одного з класів.

Головним інструментом для доведення граничних результатів, наведених у роботі, став апарат точкових процесів, за допомогою якого було одержано ряд важливих граничних теорем, пов'язаних з першими моментами надходжень $(r + 1)$ -их представників різних класів та різноманітні узагальнення цих результатів. Центральними методами у доведенні основної теореми цієї роботи стали метод пуассонізації та депуассонізації. А також було здійснено імітаційне моделювання в середовищі RStudio, за допомогою якого вдалося перевірити та візуалізувати отримані твердження.

Дана тематика й надалі буде актуальною у зв'язку з сучасною тенденцією дослідження асимптотичних властивостей дискретних імовірнісних моделей за допомогою методів теорії точкових процесів, а також з широким використанням самої задачі у різних предметних сферах, зокрема, в інформатиці, криптографії, генетиці тощо.

Список використаних джерел

- [1] Ph. Flajolet and R. Sedgewick. Analytic combinatorics. Cambridge: Cambridge University Press, 2009.
- [2] Marco Ferrante, Monica Saltalamacchia. The Coupon Collector's Problem. MATerials MATemàtics, 2014.
- [3] A. Ilienکو. Convergence of point processes associated with coupon collector's and Dixie cup problems. Electron. Commun. Probab., 2019.
- [4] P. Erdos and A. Rényi. On a classical problem of probability theory. Publ. Math. Inst. Hung. Acad. Sci., Ser. A, 6:215-220, 1961.
- [5] G. Last and M. Penrose. Lectures on the Poisson process. Cambridge: Cambridge University Press, 2017.
- [6] L. E. Baum and P. Billingsley. Asymptotic distributions for the coupon collector's problem. Ann. Math. Statist., 36(6):1835-1839, 1965.
- [7] S. I. Resnick. Extreme values, regular variation, and point processes. New York etc.:Springer-Verlag, 1987.
- [8] S. I. Resnick. Heavy-tail phenomena. Probabilistic and statistical modeling. New York, NY: Springer, 2007.
- [9] L. Holst. On sequential occupancy problems. J. Appl. Probab., 18:435-442, 1981.
- [10] O. Kallenberg. Random measures, theory and applications. Cham: Springer, 2017.
- [11] M. Penrose. Random geometric graphs. Oxford: Oxford University Press, 2003.
- [12] M. Mayer and I. Molchanov. Limit theorems for the diameter of a random sample in the unit ball. Extremes, 10(3):129-150, 2007.
- [13] E. Samuel-Cahn. Asymptotic distributions for occupancy and waiting time problems with positive probability of falling through the cells. Ann. Probab., 2(3):515-521, 1974.
- [14] M. S. Klamkin and D. J. Newman. Extensions of the birthday surprise. Journal of Combinatorial Theory 3, 279–282, 1967.

- [15] L. Holst. On birthday, collectors', occupancy and other classical urn problems. *Int. Stat. Rev.*, 54:15–27, 1986.
- [16] L. Glavaš, P. Mladenović. Limit results related to the coupon collector's problem. *Sci. Math. Hung.* Vol. 55. No. 1, 115–140, 2018.
- [17] Ph. Flajolet, P. J. Grabner, P. Kirschenhofer, H. Prodinger. On Ramanujan's Q –function. *J. Comput. Appl. Math.* Vol. 58. No. 1, 103–116, 1995.