НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Фізико-математичний факультет

Кафедра математичного аналізу та теорії ймовірності

«На правах рукопису»

УДК 519.21

До захисту допущено:

Завідувач кафедри

О.І. Клесов

«8» ____січня_2024 р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-професійною програмою «Страхова та фінансова математика»

зі спеціальності 111 «Математика»

на тему: «Центральна гранична теорема для рекордів»

Виконав :

студент II курсу магістратури, групи ОМ-21мп

Матлащук Владислав Володимирович ___________

Науковий керівник:

Професор, доктор фізико-математичних наук

Клесов Олег Іванович ___________

Рецензент

Кандидат фізико-математичних наук, доцент

Чорней Руслан Костянтинович ___________

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

Студент ______________

Київ – 2024 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Фізико-математичний факультет

Кафедра математичного аналізу та теорії ймовірності

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

Спеціальність – 111 «Математика»

Освітньо-наукова програма «Страхова та фінансова математика»

ЗАТВЕРДЖУЮ

Завідувач кафедри

Олег Іванович Клесов

Завдання

на магістерську дисертацію студенту

Матлащук Владислав Володимирович

1. Тема дисертації «Центральна гранична теорема для рекордів», науковий керівник дисертації доктор фізико-математичних наук, професор Клесов Олег Іванович, затверджені наказом по університету від « 13 » листопада 2023 р. № 5250-с
2. Термін подання студентом дисертації 11.01.2023
3. Об'єктом дослідження є Центральна гранична теорема для рекордів
4. Предметом дослідження є використання центральної граничної теореми для аналізу рекордних значень. Робота демонструє значущість цієї теореми у статистичному прогнозуванні та емпіричному аналізі екстремальних даних, особливо в спортивній статистиці.
5. Перелік завдань, які потрібно виконати:
   1) Розробка та формалізація математичної моделі для аналізу рекордів у контексті Центральної граничної теореми Ляпунова.

2) Доведення ключових властивостей Центральної граничної теореми за умови, що випадкові величини розподілені за схемою Бернуллі.

3) Аналіз умов збіжності ряду в контексті випадкових величин і доведення необхідних і достатніх умов для виконання Центральної граничної теореми.

4) Розгляд та аналіз індикаторів рекордів як випадкових величин Бернуллі.

5) Застосування отриманих теоретичних результатів до практичних прикладів і задач, де можливе виникнення рекордів.

6. Дата видачі завдання 07.09.2023

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Збір та аналіз літератури по темі дослідження. Формулювання основних гіпотез та завдань дослідження.	07.09.2023 - 21.09.2023	виконано
2	Розробка математичної моделі. Визначення методів аналізу та доведення теорем.	22.09.2023 - 13.10.2023	виконано
3	Доведення теорем 1 та 2. Аналіз умов збіжності для рекордів.	14.10.2023 - 04.11.2023	виконано
4	Доведення теорем 3 та 4.	05.11.2023 - 25.11.2023	виконано
5	Аналіз отриманих даних.	26.11.2023 - 16.12.2023	виконано
6	Написання основного тексту дисертації. Редагування та коректура.	17.12.2023 - 02.01.2024	виконано
7	Формування висновків. Підготовка презентації за результатами дослідження.	02.01.2024-08.01.2024	виконано

Студент Владислав МАТЛАЩУК

Науковий керівник Олег КЛЕСОВ

Реферат

Магістерська дисертація: 33 сторінок, 25 слайдів для проектора, 11 першоджерело.

У цьому рефераті ми досліджуємо застосування центральної граничної теореми для аналізу рекордних значень. Зокрема, вивчається статистична поведінка екстремальних значень у великих вибірках та їх вплив на прогнозування рекордів у різних сферах, включаючи спорт та інші області дослідження. Робота включає теоретичний аналіз моделей рекордних значень, заснованих на роботі видатних математиків, і практичне їх застосування для визначення ймовірності встановлення нових світових рекордів.

Мета роботи полягає в підвищення точності статистичних прогнозів та кращого розуміння динаміки рекордів.

Завданням роботи є аналіз та узагальнення статистичних методів для вивчення рекордних значень, особливо у контексті центральної граничної теореми.

Ключові слова: центральна гранична теорема, рекордні значення, екстремальні значення, статистичні методи аналізу, прогнозування рекордів, незалежні випадкові величини, розподіл ймовірностей, нормальний розподіл.

Master's thesis: 33 pages, 25 slides for projector, 11 primary sources.

In this essay, we investigate the application of the central limit theorem to the analysis of record values. In particular, we study the statistical behavior of extreme values in large samples and their impact on record prediction in various fields, including sports and other areas of research. The work includes a theoretical analysis of record models based on the work of prominent mathematicians and their practical application to determine the probability of setting new world records.

The aim of the work is to improve the accuracy of statistical forecasts and better understand the dynamics of records.

The task of the paper is to analyze and generalize statistical methods for studying record values, especially in the context of the central limit theorem.

Keywords: central limit theorem, record values, extreme values, statistical methods of analysis, record forecasting, independent random variables, probability distribution, normal distribution.

Зміст

Вступ

Вступне слово

В сучасному світі статистичний аналіз даних відіграє дуже важливу роль в багатьох різноманітних галузях від економіки та астрономії до генетики та політики. Хоча за останнє століття розвиток даного питання досяг колосального прогресу, перед людьми з кожним днем постають все тяжчі питання, які потребують покращення теперішніх знань в цій галузі або взагалі розробці цілком нових методів аналізу.

Особливе місце в статистичному аналізі посідає аналіз рекордів. Це специфічний клас даних який можна зустріти в багатьох різних галузях.

Актуальність теми

В сучасному світі ми все частіше стикаємося з цією темою, навіть не помічаючи цього. Наприклад:

- **Страхування**: компанії, які працюють у галузі страхування, щоденно використовують Центральну граничну теорему для аналізу ризиків незвичних або рідкісних подій, таких, як великі аварії, або природні катастрофи.
- **Медицина**: саме ця сфера в останні роки привернула до себе найбільше уваги, чого лише коштував всім відомий covid-19. Тому, зараз для запобігання, або контролю над росповсюджуванням захворювань використовують ЦГТ. Застосування статистичних методів дозволяє науковцям виявляти закономірності та передбачати майбутні епідемії.
- **Фінанси:** у фінансовій сфері ЦГТ використовується для оцінки потенційних ризиків. Банкіри одні з найперших, хто успішно почали використовувати Центральну граничну теорему на практиці. Яскравим прикладом може слугувати банк «Goldman Sachs», який успішно використовує ЦГТ для рекордів, чим зміг попередити декілька

глобальних криз, та зберіг гроші своїх клієнтів, заробивши звання одного з найнадійніших банків.

- **Спорт:** можна сказати, що дослідження про ЦГТ для рекордів взяло свій початок саме зі спорту, а саме для прогнозування побиття світових рекордів у різних видах спорту.
- **Клімат:** це саме та сфера, з якою у людства буде ще багато різних викликів та питань. І тут ЦГТ для рекордів стане «хорошим товаришем» для розв'язання різноманітних проблем. Навіть зараз Центральну граничну теорему використовують для досліджень рекордно високих та низьких температур, ймовірності опадів, або попередження природних аномалій

Це далеко не весь список галузей які активно використовують Центральну граничну теорему для рекордів, і з кожним роком цей список росте, роблячи цю тему дуже актуальною. Я сподіваюся внести свій невеличкий вклад в розвиток цієї теми.

Проблематика

Попри на великий спектр застосувань Центральної граничної теореми для рекордів, ця тема має багато складних аспектів та викликів.

- **Теоретичні складності:** сама Центральна гранична теорема є фундаментальною частиною теорії ймовірності, проте застосування до рекордів потребує хорошого розуміння статистичних процесів. Оскільки поведінка рекордів дуже часто відрізняється від звичайних статистичних результатів.
- **Емпіричні дані:** Отримані результати за допомогою Центральної граничної теореми, а особливо в контексті рекордів, можуть бути неповними, штучними, або мати зміщення. Це ставить під сумнів чистоту отриманих результатів. Для подолання цієї проблеми потрібна розробка складних методів для очищення та аналізу отриманих результатів.

- **Моделювання рекордів:** Використання Центральної граничної теореми для рекордів часто супроводжується моделюванням цих екстремальних подій, що своєю чергою потребує розробки спеціалізованих математичних моделей, які зможуть досить точно описати поведінку на кінцях розподілу.
- **Застосування у різних галузях:** Створення універсального методу застосування Центральної граничної теореми для різних сфер є дуже складним завдання, оскільки кожна галузь має свої унікальні характеристики, та вимоги.
- **Обчислювальні можливості:** часто для отримання досить точних результатів потрібна досить велика вибірка, що ускладнює підрахунок через потреби великих обчислювальних потужностей, та ефективних алгоритмів.

Всі ці аспекти показують, що використання Центральної граничної теореми для рекордів є досить складним завданням і вимагає глибоких теоретичних практичних знань. Проте в наш час багато проблем уже на грані вирішення, багато обов'язків бере на себе штучний інтелект, який за допомогою великих обчислювальних потужностей проводить всі потрібні підрахунки та симуляції.

Теоретичний огляд

Історія та розвиток Центральної граничної теореми

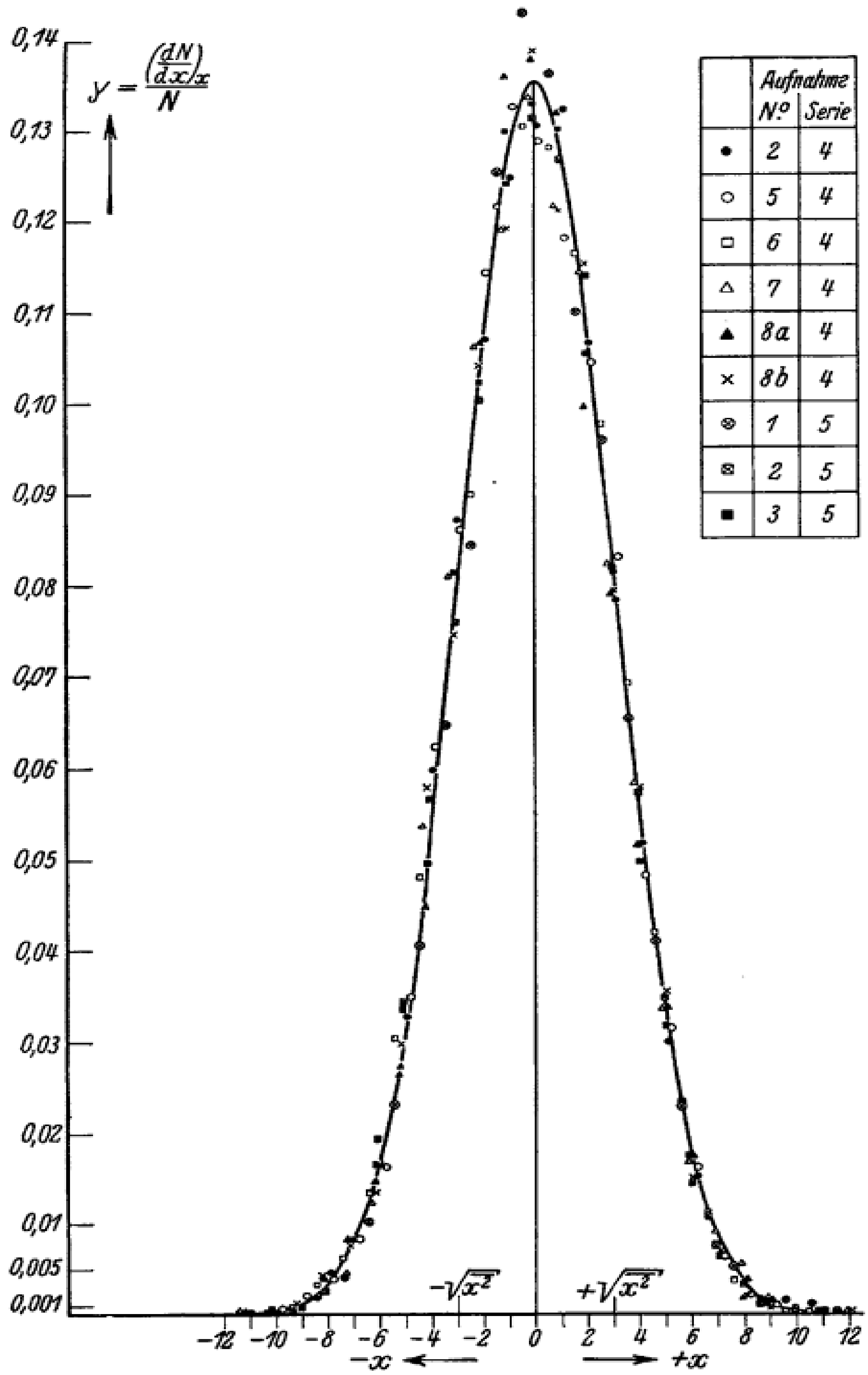
Термін «Центральної граничної теореми» бере свій початок в основному від Георга Пойя. Як він зазначив на початку своєї статті, яка було опублікована в 1920 році, «відомо, що гаусівська ймовірнісна щільність e^{-x^2}, яка зустрічається в багатьох ситуаціях, може бути пояснена однією центральною граничною теоремою», яка відіграє ключову роль у теорії ймовірності [Пойя 1920,171].

Лаплас відкрив основи цієї фундаментальної теореми у 1810 році та з позначенням «центральна гранична теорема теорії ймовірності», яке було навіть підкреслене в назві статті, Пойя дав їй назву, яка відтоді стала загальновживаною.

В наші дні «Центральна гранична теорема» пов’язана з безліччю тверджень, що стосуються збіжності ймовірнісних розподілів функцій зростаючої кількості одно- або багатовимірних випадкових змінних, або навіть більш загальних випадкових елементів до нормального розподілу. У спробі зменшити двозначність, позначення «Центральна гранична теорема» буде зазвичай відноситися до «класичного» випадку, який стосується асимптотичної рівності розподілів сум незалежних або слабо залежних випадкових змінних і не має обмеження щодо розміру. Відповідно, до даного визначення, яке відповідає поглядам Пойя «Центральна гранична теорема» насправді є збірною назвою для цілої групи теорем про збіжність функцій розподілу, щільності або дискретних ймовірностей сум випадкових величин. Одна з теорем цієї групи – яка, за загальним визначенням, була лише окремим випадком – була виведена де Муавром ще у 1733 році.

Загалом, не слід необачно застосовувати термін «Центральна гранична теорема» для будь-якого результату, пов’язаного з сумами незалежних випадкових змінних, а швидше, потрібно звертатися до центральної граничної теореми, яка розглядається на основі конкретного випадку. Проте, після того, як близько 1810 року, Лаплас використав нормальний розподіл для представлення наближень, які були справедливими в досить загальних ситуаціях, твердження що до універсального існування «закону частоти» подібного до e^{-x^2} для сум, великої кількості незалежних випадкових величин, набуло статусу природного закону в очах майже всіх ймовірнисників 19-го століття.

Цей закон послужить лейтмотивом для теорії помилок і статистики розподілу яка почалася з Кетле. Тому не дивно, що Пойя, який в певному сенсі був вкорінений у цю традицію 19-го століття, говорив про центральну граничну теорему. Розвиток цієї теореми з базової ідеї в природничих і соціальних науках в автономну математичну теорему, або, точніше, у цілу групу таких теорем є важливим етапом розвитку всієї математики.



На малюнку представлено емпіричну частотну криву відхилень крутильних вагів відповідно до їх броунівського руху [Каплер 1931]

Відхилення в певний момент часу можна вважати викликаним великою кількістю молекулярних ударів протягом певного інтервалу часу безпосередньо перед цим; отже, нормальний розподіл можна інтерпретувати як наслідок центральної граничної теореми.

Різні версії Центральної граничної теореми

В період з 1810 по 1935 роки, теореми граничних розподілів для сум (кінцево вимірних) випадкових векторів також обговорювалися час від часу, але основна увага приділялася теоремам граничних розподілів для сум незалежних і слабо залежних одновимірних випадкових величин. Сьогодні розрізняються центральні граничні теореми для нормованих сум і для трикутних масивів з одного боку, та між інтегральними та локальними теоремами граничних розподілів з іншого.

Найважливіша з історичної точки зору версія центральної граничної теореми (далі скорочено ЦГТ) є та, що належить до інтегральної теореми граничних розподілів для нормованих сум: нехай X_k буде послідовністю незалежних (або слабо залежних) випадкових величин на спільному ймовірнісному просторі. За певних умов на X_k, існують такі послідовності $a_n > 0$, що

$$\forall r \in \mathbb{R}: P\left(\frac{\sum_{k=1}^{n}(X_k - b_k)}{a_n} \le r\right) \to \Phi(r) \ \ (n \to \infty) \qquad \textit{(1.1)}$$

Φ – це функція розподілу стандартного нормального розподілу. У випадку коли

$$a_n^2 = Var\sum_{k=1}^{n} X_k \qquad i \qquad b_k = EX_k$$

говорять про «класичну нормалізацію»

Відповідні локальні граничні теореми виникають з представлення

$$P\left(\frac{\sum_{k=1}^{n}(X_k - b_k)}{a_n} \le x\right) = \int_{t \le x} f_n(t)\mu_n(t)$$

За допомогою відповідної послідовності $\sigma -$ скінченних мір (μ_n), які визначені на системі дійсних борелівських множин, та послідовності невід’ємних функцій (f_n) встановлюються умови, за яких точкова або рівномірна збіжність (f_n) до функції щільності стандартного нормального розподілу. Важливі особливі випадки включають:

- X_k є неперервною випадковою замінною, μ_n є мірою Лебега для кожного n і f_n є функцією щільності, пов'язаною з $\frac{\sum_{k=1}^{n}(X_k - b_k)}{a_n}$
- X_k набуває значень тільки $a + ih$ (де $i = 0, \pm1, \pm2, \dots$)

$$f_n(x) \coloneqq P\left(\sum_{k=1}^{n} X_k = na + zh\right)\frac{a_n}{h}$$

Для $x \in \left[\frac{na+(z-0{,}5)h-\sum_{k=1}^{n} b_k}{a_n}; \frac{na+(z+0{,}5)h-\sum_{k=1}^{n} b_k}{a_n}\right]$ $(z \in \mathbb{Z})$

і μ_n є дискретною мірою, яка присвоює вагу $\frac{h}{a_n}$ точці $\frac{na+zh-\sum_{k=1}^{n} b_k}{a_n}$

У випадку Центральної граничної теореми для трикутного масиву, розглядається подвійна послідовність випадкових змінних

$$\begin{array}{l} Y_{11} \\ Y_{21}, Y_{22} \\ Y_{31}, Y_{32}, Y_{33} \\ Y_{41}, Y_{42}, Y_{43}, Y_{44} \\ \dots \end{array}$$

розглядається, або більш загально

$$\begin{array}{l} Y_{11}, \dots, Y_{1m_1} \\ Y_{21}, \dots \dots, Y_{2m_2} \\ Y_{31}, \dots \dots \dots, Y_{3m_3} \\ Y_{41}, \dots \dots \dots \dots, Y_{4m_4} \end{array}$$

з $m_n \to \infty$. В межах кожного ряду вважається, що випадкові змінні є незалежними або слабо залежними. Інтегральна форма ЦГТ досліджує збіжність розподілів сум рядків, до нормального розподілу, тобто визначає, за яких умов вірне наступне

$$\forall r \in \mathbb{R}: P\left(\sum_{k=1}^{m_n} Y_{nk} \le r\right) \to \Phi(r) \qquad (n \to \infty) \qquad (1.2)$$

Так само як і для нормованих сум, локальні версії ЦГТ можуть бути розглянуті для трикутних масивів. Очевидно, за визначенням $Y_{nk} = \frac{X_k - b_k}{a_n}$ ЦГТ для нормованих сум (1.1) стає окремим випадком ЦГТ для трикутних масивів (1.2) якщо вважати заданими нормуючі константи a_n та b_n. Однак, проблема нормованих сум також ставить питання про те, як знайти відповідні a_n та b_n.

З певними модифікаціями розглянуті задачі можна також перенести на ЦГТ для випадкових векторів.

Чебишов [1887/90] був першим хто сформулював твердження використовуючи ЦГТ (1.1) для послідовності незалежних випадкових величин, використовуючи класичне нормування і він намагався довести своє твердження за певних умов. В кінці 19-го на початку 20-го століття математики в основному розглядали співвідношення еквівалентні (1.1):

$$\forall\, a \le b \in \mathbb{R} : P\left(a \le \frac{\sum_{k=1}^{n}(X_k - b_k)}{a_n} \le b\right) \to \Phi(b) - \Phi(a) \quad (n \to \infty)$$

Або

$$\forall\, a < b \in \mathbb{R} : P\left(a < \frac{\sum_{k=1}^{n}(X_k - b_k)}{a_n} < b\right) \to \Phi(b) - \Phi(a) \quad (n \to \infty)$$

Формулювання інтегральних граничних теорем саме у формі (1.1) ймовірно належить фон Мізесу [1919]. Локальні граничні теореми явно з'являються в роботі Некрасова [1898]. Звернення до некласичного нормування почалося всерйоз у 1920-х роках з [Бернстайн 1922; 1926] і [Леві 1925]. Ідею розглядів трикутних масивів можна знайти в [Бернстайн 1922] і [Ліндеберг 1922], але насправді інтенсивно її розробляли лише в 1930-х. Розгляд центральної граничної теореми для сум залежних випадкових величин розпочав [Марков 1907/10]. З моменту публікації статті Чебишова і до середини 1930-х років ЦГТ (1.1) для незалежних випадкових величин постійно перебувала в центрі уваги.

До Чебишова люди вивчали фактично не граничні теореми, а займалися приблизними обчисленнями ймовірнісних густин, індивідуальних ймовірностей або ймовірностей того, що сума випадкових змінних лежить «між» попередньо визначеними межами, в абсолютному або відносному сенсі. Відповідні твердження можна тлумачити з точки зору як нормованих сум за класичним нормуванням, так і трикутних масивів. Приблизні обчислення для густини (у випадку неперервних випадкових змінних) або дискретних ймовірностей (у випадку гранчасто розподілених випадкових змінних) та приблизні обчислення для інтегральних ймовірностей вважалися еквівалентними. З цієї причини немає великого сенсу розрізняти між інтегральними та локальними формами ЦГТ в епоху до Чебишова.

Центральні граничні теореми. Доведення

Теорема Чебишова

Нехай $X_1, X_2, X_3, \dots, X_n$ – послідовність попарно незалежних випадкових величин, які задовольняють умовам

$$1) E(X_i) = a_i$$
$$2) Var(X_i) \leq c \quad \text{для усіх } i = 1 \div n$$

Тоді

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} a_i}{n}\right| < \varepsilon\right) = 1 \qquad (2.1)$$

Доведення

Для початку знайдемо математичне сподівання та дисперсію середньої випадкової величини, тобто $\frac{X_1+X_2+X_3+\cdots+X_n}{n}$

$$E\left(\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}\right) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$Var\left(\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}\right) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \leq \frac{nc}{n^2} = \frac{c}{n}$$

Застосуємо для випадкової величини $\frac{1}{n}\sum_{i=1}^{n} X_i$ нерівність Чебишова

Нерівність Чебишова, яка є одним з основних інструментів теорії ймовірності, дає оцінку того, як велика частина ймовіностей може зосередитися на відстані від середнього значення випадкової величини.
Нехай X є випадковою величиною з математичним сподіванням $E(X)$ та дисперсією $Var(X)$. Для будь $-$ якого додатнього числа k, нерівність Чебишова стверджує, що

$$P(|X - E(X)| \geq k) \leq \frac{Var(X)}{k^2}$$

Ця нерівність каже, що ймовірність того, що значення випадкової величини відхилиться від її середнього на величину k або більше, є не більшою, ніж відношення її дисперсії до квадрату цього відхилення

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} a_i\right| < \varepsilon\right) \geq 1 - \frac{c}{n\varepsilon^2} \qquad (2.2)$$

Границя цієї ймовірності при $n \to \infty$ дорівнює одиниці, тобто рівність (2.1) доведена. ■

Наслідок 1

(Теорема Чебишова для однаково розподілених випадкових величин)

Якщо в умові теореми Чебишева випадкові величини $X_1, X_2, X_3, \dots, X_n$ мають однакове математичне сподівання $E(X_i) = a, Var(X_i) = \sigma^2$ то для будь-якого $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| < \varepsilon\right) = 1 \qquad (2.3)$$

З (2.2) отримуємо оцінку

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| < \varepsilon\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \qquad (2.4)$$

(2.3) можна записати у вигляді

$$\lim_{n\to\infty} \left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| \geq \varepsilon\right) = 0 \qquad (2.5)$$

При досить великому числі випадкових величин, таких, які мають рівні математичні сподівання, їх середнє значення відхиляється від їх математичного сподівання скільки завгодно мало по ймовірності

$$\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} \to a \qquad (2.6)$$

Перший наслідок дуже часто використовується. При наявності великої кількості випадкових величин, середнє значення утрачає силу випадкової величини. В цьому можна виявити зв’язок між необхідністю та випадковістю.

При виконанні експерименту для встановлення певного показника, як правило, за результуюче значення приймають середнє арифметичне з усіх отриманих величин.

Цей підхід є валідним, коли всі тести є взаємно незалежними, математичні очікування збігаються, а вимірювальний прилад забезпечує достатній рівень точності.

Наслідок 2 (теорема Бернуллі)

Нехай ймовірність появи події A в кожному з n незалежних повторних випробувань дорівнює p, m – число появ події A (частота подій) в n випробуваннях. Тоді

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1 \qquad \varepsilon > 0$$

Доведення

Частоту $\frac{m}{n}$ можна розглянути як невід’ємну випадкову величину X. Знайдемо математичне сподівання

$$E\left(\frac{m}{n}\right) = \frac{1}{n}E(m) = \frac{1}{n}np = p$$

Потрібно оцінити ймовірність відхилення випадкової величини $\frac{m}{n}$ від математичного сподівання p. Для цього знайдемо дисперсію

$$Var\left(\frac{m}{n}\right) = \frac{1}{n^2}Var(m) = \frac{1}{n^2}npq = \frac{pq}{n} = \frac{p(1-p)}{n}$$

За нерівністю Чебишова отримуємо

$$P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \geq 1 - \frac{p(1-p)}{n\varepsilon^2} \qquad (2.8)$$

Застосувавши граничний перехід при $n \to \infty$ отримуємо (2.7) що і треба було довести.■

Центральна гранична теорема

Нехай задана послідовність незалежно однакових розподілених величин $X_1, X_2, X_3, \dots, X_n$

$$E(x_i) = 0$$
$$Var(x_i) = b \qquad i = 1 \div n$$

Розглянемо випадкову величину $A_n = \sum_{i=1}^{n} X_i$

Тоді $E(A_n) = \sum_{i=1}^{n} E(X_i) = 0$

$$Var(A_n) = \sum_{i=1}^{n} Var(X_i) = bn$$

при $n \to \infty$ функція розподілу $F_{A_n}(x) = P(A_n < x) \to \frac{1}{\sqrt{2\pi nb}} \int_{-\infty}^{x} e^{-\frac{z^2}{2nb}} dz$, тобто сума A_n буде розподілена за нормальним законом з математичним сподіванням, що дорівнює нулеві та дисперсією nb.

Наслідок

Коли кількість $n \geq 30$ розподіл суми однаково розподілених випадкових величин мало чим відрізняється від нормального розподілу.

Теорема Ляпунова

Нехай маємо послідовність незалежних випадкових величин

$X_1, X_2, X_3, \dots, X_n, \dots$ таких, що $E(x_i) = 0$

$\mathrm{Var}(\mathrm{X_i}) = \mathrm{b_i^2}, \quad \mathrm{i} = 1{,}2{,}3, \dots \mathrm{n}, \dots$

Побудуємо суму випадкових величин $A_n = \sum_{i=1}^{n} X_i$. Позначимо $B_n^2 = \sum_{i=1}^{n} b_i^n$, якщо виконується умова рівної малості величин, що утворює суму

$$\frac{1}{B_n^3}\sum_{i=1}^{n} E(X_i)^3 \to 0 \quad \text{при } n \to \infty$$

то сума A_n буде нормально розподілена з математичним сподіванням

$E(A_n) = 0 \quad$ та $\quad Var(A_n) = B_n^2$

Незалежно від закону розподілу окремих незалежних доданків, кожен з яких має обмежене середнє та дисперсію, розподіл їх сум, коли вона нормується, асимптотично прямує до нормального розподілу при $n \to \infty$.

Цей значний аспект залишається справедливим навіть тоді, коли окремі доданки мають різний розподіл, доки внесок кожної незалежної випадкової величини $X_1, X_2, X_3, \dots, X_n$ залишається незначним у загальній сумі при

$n \to \infty$.

Максимуми та рекорди

Означення

Нехай $X_1, X_2, \dots, X_n$ є послідовністю незалежних однаково розподілених випадкових величин з неперервною функцією розподілу $F(x)$.Тоді значення X_i є "рекордом" у цій послідовності, якщо воно перевищує всі попередні спостереження. Формально, рекорд на позиції i визначається так:

$R_i = X_i$ є рекордом, якщо $X_i > \max\{X_1, X_2, \dots, X_{i-1}\}$.

Розглянемо випадкову вибірку $X_1, X_2, X_3, \dots, X_n$ розміру n із популяції з неперервною функцією розподілу F. Вілкс (1959) знайшов розподіл мінімальної кількості додаткових спостережень $N(n)$, які необхідні, щоб отримати спостережуване значення, яке перевищує $M_n = \max(X_1, \dots, X_n)$. Легко помітити, що

$$P\{N(n) > m\} = P\{max\{X_{n+1}, \dots, X_{n+m}\} \leq M_n\} \qquad (3.1)$$

$$= \int_{-\infty}^{\infty} F^m(u) d\big(F^n(u)\big) = n \int_{-\infty}^{\infty} F^{m+n-1}(u) dF(u)$$

$$= n \int_0^1 v^{m+n-1} dv = \frac{n}{n+m}$$

і тоді

$$E\,N(n) = \sum_{m=0}^{\infty} P\{N(n) > m\} = \sum_{m=0}^{\infty} \frac{n}{m+n} = n \sum_{k=n}^{\infty} \frac{1}{k} = \infty \qquad (3.2)$$

Права частина формули (3.1) наводить нас на додатковий (досить елементарний) доказ цієї рівності. Дійсно, завдяки неперервності F незалежні випадкові змінні $X_1, \dots, X_{n+m}$ набувають майже напевно різних значень, і оскільки всі вони мають однаковий розподіл, подія $\{M_{n+m} = X_k\}$,

$1 \leq k \leq n+m$ є рівноймовірним

$$P\{M_{n+m} = X_k\} = \frac{1}{n+m}$$

Залишається зауважити, що

$$P\{max\{X_{n+1}, \dots, X_{n+m}\} \le M_n\} = P\{M_{n+m} = M_n\}$$
$$= \sum_{k=1}^{n} P\{M_{n+m} = X_k\} = \frac{n}{n+m}$$

За симетрією ми отримаємо той самий розподіл, як і (3.1), якщо розглядатимемо додаткову кількість спостережень, необхідних для отримання спостереження, яке потрапляє в інтервал $(-\infty, m_n)$, де

$m_n = \min\{X_1, \dots, X_n\}$. Тепер нехай $N(1,n)$ позначає мінімальний розмір додаткової вибірки, що містить хоча б одне спостереження, що виходить за межі інтервалу $[\mathrm{m_n}, M_n]$. Тоді

$$P\{N(1,n) > m\} = P\{m_n \le X_{n+k} < M_n, k = 1,2,3, \dots, m\}$$

$$= \int_{-\infty}^{\infty} \int_{u}^{\infty} \big(F(v) - F(u)\big)^m n(n-1)\big(F(v) - F(u)\big)^{n-2} dF(u)dF(v) = \qquad (3.3)$$

$$= n(n-1) \int_{-\infty}^{\infty} \int_{u}^{\infty} \big(F(v) - F(u)\big)^{n+m-2} dF(u)dF(v) = \frac{n(n-1)}{(n+1)(n+m-1)}$$

$$\times \int_{-\infty}^{\infty} \int_{u}^{\infty} (n+m)(n+m-1)\big(F(v) - F(u)\big)^{n+m-2} dF(u)dF(v) =$$

$$= \frac{n(n-1)}{(n+m)(n+m-1)} P\{-\infty < m_{n+m} < M_{n+m} < \infty\} = \frac{n(n-1)}{(n+m)(n+m-1)}$$

І

$$E\,N(1,n) = \sum_{m=0}^{\infty} P\{N(n) > m\} = n(n-1) \sum_{m=0}^{\infty} \frac{1}{(n+m-1)(n+m)} \qquad (3.4)$$

$$= n(n-1) \sum_{m=n}^{\infty} \frac{1}{m(m-1)} = n(n-1) \sum_{m=n}^{\infty} \left(\frac{1}{m-1} - \frac{1}{m}\right) = n$$

Як і у випадку (13.1), відношення (13.3) також допускає простий випадок. В цьому випадку задачу можна звести до наступного: з $n + m$ місць випадковим чином вибираються два для присвоєння максимуму та мінімуму на них, і ми повинні знайти ймовірність, що числа в обох місцях не перевищують n. Простий розрахунок дає для цієї ймовірності значення

$$\frac{n(n-1)}{(n+m)(n+m-1)}$$

Часові рекорди у F^{α}- схемі

Означення. Послідовність незалежних випадкових величин $X_1, X_2, \ldots$ з неперервними функціями розподілу $F_1, F_2, \ldots$ утворює $F^{\alpha} -$ схему, якщо $F_k = F^{\alpha(k)}, \quad k = 1,2, \ldots$, де F є неперервною функцією розподілу, а $\alpha(1), \alpha(2), \ldots$ є деякими додатними константами.

Модель Янга: Моделі з трендом у вигляді $Y_n = X_n + c(n), n = 1,2,3, \ldots$ включають незалежні, однаково розподілені базові випадкові змінні $X_1, X_2, \ldots$. . У моделі Пфайфера випадкові змінні в кожній серії між двома послідовними рекордами мають ідентичний розподіл.

- **Модель Пфайфера**. модель записів яка напряму пов'язана зі спортом. Німецький легкоатлет Уве Хон встановив рекорд у метанні списа 20 липня 1984 року, прокинувши спис на відстань 104.8 метра, що стало небезпечним для глядачів. Тоді конструкцію списа було змінено, зсунувши центр тяжіння, що змінило динаміку польоту списа. Рекорди з новим списом ще не перевищили показник Хона. Але найкращі метання тепер наближаються до позначки 100 метрів, і без сумніву правила зміняться знову, як тільки ця позначка буде перевищена. Ідея розподілів випадкових змінних у послідовності $X_1, X_2, X_3, \ldots$ є основною для схеми Пфайфера (1982,1984).
  Нехай $\{X_{n,k}, n \geq 1, k \geq 1\}$ буде подвійним масивом незалежних випадкових змінних, кожен рядок якого $X_{n,1}, X_{n,2}, \ldots$ має спільну

функцію розподілу F_n. Буде зручно визначати спершу часи між рекордами $\Delta(n) = L(n) - L(n-1)$ замість часу рекордів $L(n)$. Вони визначені як

$$\Delta(1) = 1$$

$$\Delta(n+1) = \min\{k: X_{n+1,k} > X_{n,\Delta(n)}\}, n = 1,2, \dots$$

Тоді часи рекордів $\mathrm{L}(n)$ і значення рекордів $X(n)$ у схемі Пфайфера визначаються як $L(n) = \Delta(1) + \dots + \Delta(n)$

$$X(n) = X_{n,\Delta(n)}, \quad n = 1,2, \dots$$

Якщо $F_1 = F_2 = \cdots$ то ці визначення збігаються з визначеннями класичних статистик записів.

У моделі Балабекян-Невзоров максимальні результати в кожному циклі

- **Схема Балабекяна–Невзорова**. Ще одна схема записів була запропонована Балабекяном та Невзоровим (1986) та розвинена Ранненом (1991). Розгляньмо знову спортивні змагання, наприклад, стрибки у довжину. Нехай m спортсменів різної майстерності здійснюють по n спроб кожен. Ми можемо припустити, що їхні довжини стрибків описуються послідовністю незалежних випадкових змінних

$$X_1, \dots, X_m, \dots X_{m(n-1)+1}, \dots, X_{nm}$$

$$\text{з функціями розподілу}$$

$$F_1, \dots, F_m, \dots, F_{m(n-1)+1}, \dots, F_{mn}$$

$$\text{таким чином, що}$$

$$F_{mk-(m-1)+r} = F_r, \quad k = 1, \dots, n, \quad r = 1,2, \dots, m$$

тобто, ця послідовність складається з n повторень групи з m функцій розподілу $F_1, \dots, F_m$. Ця схема поєднує особливості ненормальної моделі (m різних функцій розподілу) та класичної моделі записів, оскільки найбільші стрибки в кожній спробі

$$Y_k = \max\{X_{m(k-1)+1}, \dots, X_{mk}\}, \quad k = 1,2, \dots, n$$

формують послідовність незалежних однаково розподілених випадкових змінних зі спільною функцією розподілу.

$$G = \prod_{r=1}^{m} F_r$$

Отже, ми розглядаємо послідовність незалежних випадкових змінних $X_1, X_2, \dots$ з неперервними функціями розподілу $F_1, F_2, \dots$ такими, що

$$F_{mk-(m-1)+r} = F_r, \;\; r = 1,2,\dots,m \quad k = 1,2,\dots$$

Нехай $N(nm)$ позначає кількість записів у послідовність

$$X_1, \dots, X_m, \dots, X_{m(n-1)+1}, \dots, X_{mn}$$

тобто, кількість записів, підрахованих після n пробігів, кожен з яких складається з m випадкових змінних.

$Y_k = \max\{X_{m(k)+1}, \dots, X_{m_k}\}, \;\; k = 1,2,\dots$

утворюють стаціонарну послідовність. Перша модель, яку можна було б розглядати як справді незалежну, була запропонована Янгом (1975). Аналізуючи динаміку Олімпійських рекордів, Янг виявив, що періоди часу між встановленням наступних рекордів не суперечать інтер-рекордним часам в класичній моделі рекордів. Він припустив, що найкращий результат, показаний на деяких змаганнях Олімпійських ігор, можна трактувати як найкращий результат у цьому виді спорту серед всього світового населення. Таким чином, Янг запропонував розглядати рекорди в наступній послідовності

$Y_k = \max\{X_{k,1}, \dots, X_{k,n(k)}\}, \quad k = 1,2,\dots$

де $\{X_{i,j}\}, \quad j = 1,2,3,\dots.,n(i), \quad i = 1,2,\dots$

є незалежними однаково розподіленими випадковими змінними з загальною безперервною функцією розподілу F залежно від події,$n(k)$ є розміром населення світу на момент k-тих Олімпійських ігор. Числа

$n(k) = \lambda^{k-1} n(1), \quad k = 1,2, \dots$

Коефіцієнт λ, який представляє геометричне зростання населення нашої планети, був обраний як $2^{\frac{1}{9}} \approx 1.08$ на основі того факту, що протягом дев'яти чотирирічних періодів між послідовними Олімпійськими іграми з 1900 по 1936 роки світове населення подвоїлося. Фактично в схемі Янга розглядаються записи в послідовності незалежних випадкових змінних $Y_1, Y_2, \dots,$ з функціями розподілу

$F_k(x) = \left(F(x)\right)^{n(k)}$

де $n(k) = \lambda^{k-1} n(1), \quad k = 1,2, \dots,$ і F є неперервною функцією розподілу. Янг отримав наступні результати

$\Delta(n) = L(n) - L(n-1), \; n = 2,3,4, \dots$

Самостійне дослідження

Теорема 1

Розглянемо «Центральну граничну теорему Ляпунова» для випадку коли всі випадкові величини X_n незалежні та Бернуллі. Тобто, кожна величина приймає лише два значення:

$$P(X_n = 1) = P_n$$
$$P(X_n = 0) = 1 - P_n$$

Підрахуємо математичне сподівання та дисперсію для X_n:

$$E\, X_n = P_n$$
$$var[X_n] = E(X_n - P_N)^2 = (1 - P_n)^2 P_n + (-P_n)^2 (1 - P_n)$$
$$= (1 - P_n) P_n (1 - P_n + P_n) = P_n (1 - P_n).$$

Порахуємо 3-й момент:

$E|X_n - P_n|^3 = (1 - P_n)^3 P_n + P_n^3 (1 - P_n) = P_n (1 - P_n)((1 - P_n)^2 + P_n^2).$

Розглянемо результат який ми отримали:

$$P_n < 1 \text{ тому } P_n^2 < 1$$

$$1 - P_n < 1 \text{ і також } (1 - P_n)^2 < 1, \text{отже}$$

$$((1 - P_n)^2 + P_n^2) < 2.$$

В результаті ми отримали що: $P_n(1 - P_n)((1 - P_n)^2 + P_n^2) < 2P_n(1 - P_n) = 2var[X_n]$.

Тепер застосуємо отримані дані для «рекордів в $F^{\alpha} -$ схемах», для цього потрібно використати «Індикатори рекорді»:

$$I_n - \text{ідикатор, де } n - \text{момент рекорду}$$

Якщо це індикатор, то він може приймати лише два значення:

0 – немає рекорду

1 – є рекорд.

Тобто можна сказати, що $I_n -$ це випадкова величина Бернуллі.

Тепер можемо розглянути таку ймовірність (за теоремою Невзорова):

$$P(I_n = 1) = \frac{\alpha_n}{A_n},$$

$$\alpha_n - \text{це число з } F^{\alpha} - \text{схеми } \{\alpha_n\},$$
$$A_n = \alpha_1 + \alpha_2 + \cdots + \alpha_n,$$

$$\mu_n - \text{кількість рекордів до } n,$$
$$\mu_n = I_1 + I_2 + \cdots + I_n.$$

Запишемо умови Ляпунова:

$$S_n^2 = \sum_{i=1}^{n} \sigma_i^2, \qquad \sigma_i^2 = var[x_i] \ ,$$

$$\sum_{i=1}^{n} E\ |X_i - P_i|^3.$$

І якщо,

$$\frac{\sum_{i=1}^{n} E|X_i - P_i|^3}{S_n^3} \to 0,$$

то виконується Центральна гранична теорема

Перевіримо для Бернуллі, ми з’ясували, що третій момент для Бернуллі не перевищує дві дисперсії:

$$\frac{\sum_{i=1}^{n} E|X_i - P_i|^3}{S_n^3} \le \frac{\sum_{i=1}^{n} 2\sigma_i^2}{S_n^3},$$

$$\frac{\sum_{i=1}^{n} 2\sigma_i^2}{S_n^3} = \frac{2S_n^2}{S_n^3} \xrightarrow{S_n^2 \to \infty} 0.$$

$$S_n^2 = \sum_{i=1}^{n} \sigma_i^2 = \sum_{i=1}^{n} P_i(1 - P_i) = \sum_{i=1}^{n} \frac{\alpha_i}{A_i}\left(1 - \frac{\alpha_i}{A_i}\right) - \text{загальний варіант.}$$

Отже, для «рекордів» умова Ляпунова зводиться до

$$\sum_{i=1}^{n} \frac{\alpha_i}{A_i}\left(1 - \frac{\alpha_i}{A_i}\right) - \text{розбігається.}$$

Припустимо, що

$$\frac{\alpha_i}{A_i} \to 0.$$

Тоді, $\left(1 - \frac{\alpha_i}{A_i}\right) \approx 1.$

Тепер можемо сказати що умова розбіжного ряду еквівалентна тому, що розбігається такий ряд:

$$\sum_{i=1}^{n} \frac{\alpha_i}{A_i} - \text{розбігається.}$$

Тобто, якщо у нас виконуються умови:

1) $\frac{\alpha_i}{A_i} \to 0$
2) $\sum_{i=1}^{n} \frac{\alpha_i}{A_i} - \text{розбігається}$ Спрощений варіант

То виконується умова Ляпунова і виконується «Центральна гранична теорема» для кількості рекордів.

Теорема 2

Нехай маємо F^{α} − схему, де $\alpha(j) = 1$ або $\alpha(j) = 2.$ Тоді виконується Центральна гранична теорема.

Доведення

$$A_n = \{1,3,4,6,7,\dots\} = \sum_{j=1}^{n} \frac{\alpha(j)}{\alpha(1) + \alpha(2) + \cdots + \alpha(j)}.$$

Розглянемо дану суму для двох випадків, коли α_j − парне, та коли α_j − не парне.

1) Для парних:

$$A_n = \sum_{m=1}^{n} \frac{2}{3m}$$

$$A_n = \frac{2}{3}\sum_{m=1}^{n} \frac{1}{m} - \text{гармонічний ряд.}$$

2) Для не парних:

$$A_n = \sum_{m=1}^{n} \frac{1}{3m-1}$$

$$\frac{1}{3m-1} = \frac{1}{3m} + \frac{1}{3m(3m-1)}$$

$$A_n = \sum_{m=1}^{n} \frac{1}{3m} + \sum_{m=1}^{n} \frac{1}{3m(3m-1)} =$$

$$= \frac{1}{3}\sum_{m=1}^{n} \frac{1}{m} + \sum_{m=1}^{n} \frac{1}{3m(3m-1)}.$$

Ми отримали $\frac{1}{3}\sum_{m=1}^{n}\frac{1}{m}$ – гармонічний ряд,

та $\sum_{m=1}^{n}\frac{1}{3m(3m-1)}$ – збіжний ряд.

Тепер запишемо формулу для загального випадку:

$$A_n = \frac{1+(-1)^n}{3}\left(\ln(n) + \gamma + \frac{1}{2n} + \frac{1}{12n^2} + \frac{1}{120n^4} + \cdots\right) + \frac{1-(-1)^n}{2}\sum_{m=1}^{n}\frac{1}{3m(3m-1)}$$

Тепер перевіримо чи виконуються умови.

Для даного випадку краще використовувати спрощений варіант:

$$1)\ \frac{\alpha_i}{A_i} \to 0$$

$$2)\ \sum_{i=1}^{n} \frac{\alpha_i}{A_i} - \text{розбігається.}$$

Оскільки у нас в знаменнику знаходиться $\ln(n)$, можна записати ,що

$$A_i \sim \ln(i).$$

Тобто перша умова виконується, оскільки α_i у нас $const$, то

$$\frac{const}{\ln(i)} \to 0.$$

Перевіримо 2 умову:

$$\sum_{i=1}^{n} \frac{const}{\ln(i)} - \text{розбігається.}$$

Тобто виконуються обидві умови, а значить виконується «Центральна гранична теорема» для кількості рекордів.

Теорема 3

Нехай F^{α} – схема така, що $2\alpha(n) = A_n$. Тоді виконується Центральна гранична теорема.

Доведення

Нашу умову можемо записати у вигляді $\frac{\alpha(n)}{A\_n} = \frac{1}{2}$.

Для даного випадку зручніше використовувати загальну умову:

$$\sum_{i=1}^{n} \frac{\alpha_i}{A_i}\left(1 - \frac{\alpha_i}{A_i}\right) - \text{розбігається.}$$

Підставляємо для нашого випадку:

$$\sum_{i=1}^{n} \frac{1}{2}\left(1 - \frac{1}{2}\right) = \sum_{i=1}^{n} \frac{1}{4}.$$

Оскільки не виконується необхідна умова збіжності, цей ряд розбігається.

Тобто для даного випадку також виконується «Центральна гранична теорема» для кількості рекордів.

Теорема 4

Нехай F^{α} – схема така, що $\frac{\alpha_n}{A_n} = \frac{1}{n^2}$. Тоді виконується Центральна гранична теорема.

Доведення

Для даного випадку зручно користуватися спрощеним варіантом:

$$1)\ \frac{\alpha_i}{A_i} \to 0$$

$$2)\ \sum_{i=1}^{n} \frac{\alpha_i}{A_i} - \text{розбігається}$$

- 1) $\frac{\alpha_i}{A_i} \to 0$ $\qquad\qquad$ $\frac{1}{n^2} \to 0$
- 2) $\sum_{i=1}^{n} \frac{\alpha_i}{A_i} -$ розбігається, $\qquad$ $\sum_{i=1}^{n} \frac{1}{i^2} -$ збігається.

Не виконується 2 умова і не виконується «Центральна гранична теорема» для кількості рекордів.

Висновок

У рамках даної магістерської дисертації було здійснено глибокий теоретичний та практичний аналіз використання центральної граничної теореми для аналізу рекордних значень. Робота виявила значущість цієї теореми в статистичному прогнозуванні та в емпіричному аналізі екстремальних даних. Встановлено, що центральна гранична теорема забезпечує надійну основу для оцінювання ймовірності рекордних подій у великих вибірках, що має велике значення у багатьох областях, зокрема в спортивній статистиці.

Дослідження підтвердило, що з розширенням обсягу даних розподіл значень схиляється до нормального, дозволяючи точніше прогнозувати екстремальні події. Це виявлення має важливе практичне застосування, особливо в контексті оцінювання рекордів та їхньої можливої частоти.

Результати дисертації не лише підтверджують наявну теорію, а й розширюють границі застосування статистичних методів аналізу, відкриваючи нові можливості для дослідників та професіоналів, які працюють з великими даними. Отримані знання можуть бути використані для подальшого розвитку моделей прогнозування в різних секторах і допоможуть формувати більш точні та обгрунтовані стратегії прийняття рішень.

У даній роботі була розглянута

Цетральна гранична теорема для $F^{\alpha} -$ схеми.

1. $F^{\alpha} -$ схема, де $\alpha(j) = 1$ або $\alpha(j) = 2$.
2. $F^{\alpha} -$ схема, де $2\alpha(n) = A_n$.
3. $F^{\alpha} -$ схема, де $\frac{\alpha_n}{A_n} = \frac{1}{n^2}$.

В першому та другому випадках Центральна гранична теорема виконується, в третьому не виконується.

Список використаної літератури

1. Nevzorov, V. B. (2000). Records: Mathematical Theory.
2. Ahsanullah, M., & Nevzorov, V. B. (2015). Records via Probability Theory.
3. Fisher, H. (2010). A History of the Central Limit Theorem.
4. A. Arnold, "Understanding the Central Limit Theorem
5. N. Balakrishnan, "Advances in Order Statistics
6. B. I. G. Barron, "The Central Limit Theorem and Its Applications
7. Laurens de Haan i Ana Ferreira "Extreme Value Theory: An Introduction"
8. William Feller "An Introduction to Probability Theory and Its Applications"
9. Emil Gumbel "The Theory of Extreme Values"
10. Michael Falk, Jürg Hüsler, i Rolf-Dieter Reiss "Records and Branching Processes"
11. Paolo Giudici, Silvia Figini "Statistical Models for Data Analysis"