

Дипломна робота

на здобуття ступеня бакалавра

на тему:

**Застосування інтегральних методів максимальної
правдоподібності для точкового оцінювання в умовах
наявності несуттєвих параметрів**

Автор: студентка 4 курсу, групи КА-21,
ННК «ІПСА» НТУУ «КПІ»
Гаврилюк Галина Ігорівна

Керівник: канд. фіз.-мат. наук, доцент
Каніовська І. Ю.

Об'єкт дослідження:

Різноманітні задачі точкового оцінювання параметрів генеральної сукупності за умов наявності несуттєвих параметрів, зокрема, задача оцінювання параметра N біноміального розподілу (фізичний сенс N – кількість шахрайських дій у платіжній системі за деякий період часу).

Предмет дослідження:

Методи максимальної правдоподібності (класичний ММП, однорідний інтегральний ММП, інтегральний ММП Джефрі, профільний та умовний ММП).

Мета:

Визначити та проілюструвати на прикладах переваги та недоліки застосування інтегральних методів максимальної правдоподібності в контексті інших методів, що використовують функції правдоподібності.

Актуальність дослідження:

Інформація –
цінний ресурс



Точкове оцінювання
параметрів



Закономірності у значеннях
випадкових величин



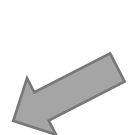
Конкурентна перевага,
нові знання



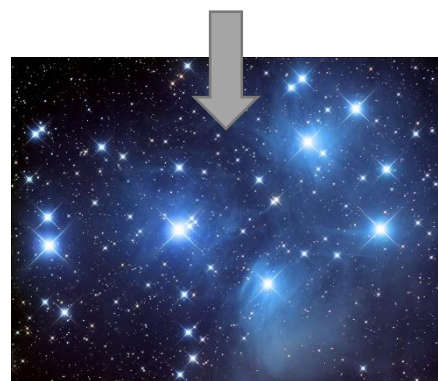
Maths



Оцінка кількості особин
деякого виду



Запропонований
підхід



Оцінка кількості зірок на небі



Оцінка ефективності
інструментів виявлення
шахрайських дій у платіжних
системах

Точкове оцінювання параметрів ГС

Точною оцінкою параметра θ є статистика $\theta^* = \theta^*(\xi_1, \xi_2, \dots, \xi_n)$, значення якої знайдено за конкретною реалізацією вибірки і береться за наближене значення невідомого параметра:

$$\theta_{\text{зн}}^* = \theta^*(x_1^0, x_2^0, \dots, x_n^0) \approx \theta$$

Вимоги до точкових оцінок:

Незміщеність

- Означення: $E\theta^* = \theta$

Конзистентність

- Означення: $\forall \varepsilon > 0: \lim_{n \rightarrow \infty} P\{|\theta_n^* - \theta| > \varepsilon\} = 0$

Критерій: $D\theta_n^* \xrightarrow{n \rightarrow \infty} 0$ (для незміщених)

Ефективність

- Означення: $\hat{D}\theta^* = \min_{\theta_j^* \in \Theta} D\theta_j^*$

Критерій: $\frac{\partial \ln L_{\xi}(\vec{\xi}, \theta)}{\partial \theta} = C(n, \theta)(\theta^*(\vec{\xi}) - \theta)$ (для незміщених)

Що таке функція правдоподібності?

Для дискретних ξ :

$$L(\vec{x}) = \prod_{k=1}^n P(\xi = x_k)$$

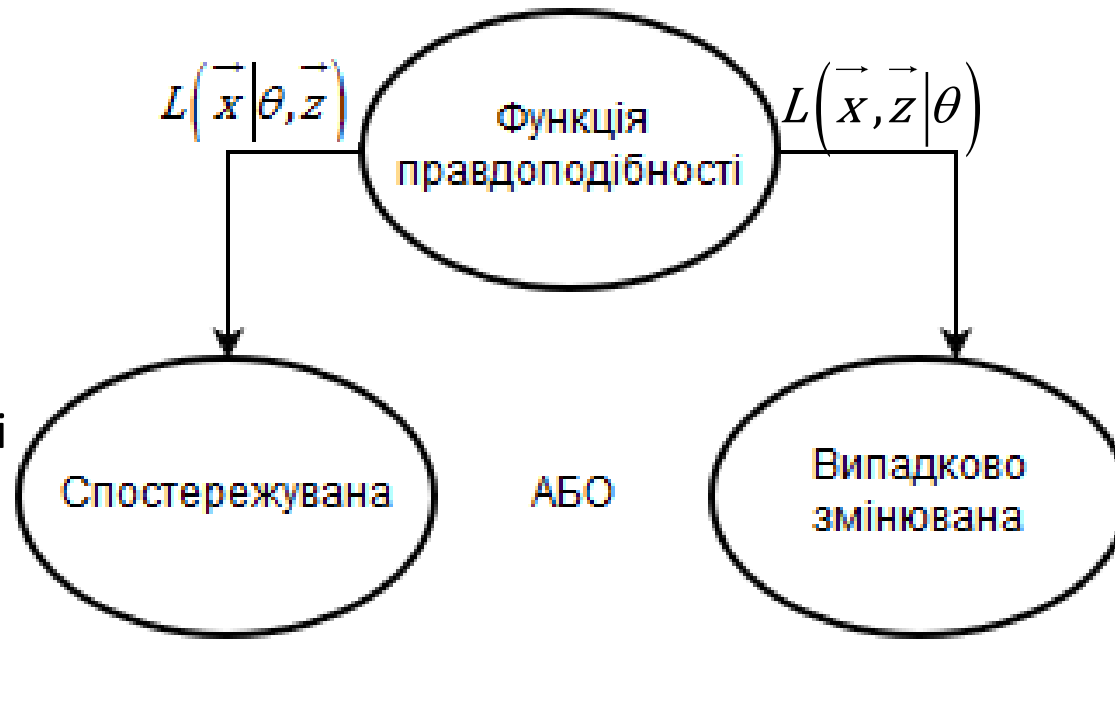
Для неперервних ξ :

$$L(\vec{x}) = \prod_{k=1}^n f_{\xi}(x_k)$$

Поняття функції правдоподібності можна пояснити на прикладі підкидання монети 10 разів.

Думка: єдиного визначення функції правдоподібності не може існувати.
Загальноприйняте визначення Саваджа (вище) не пояснює, як трактувати неспостережувані величини (наприклад, майбутні спостереження $\vec{z}$).

Рішення: розглядати ФП у вигляді: $L(\vec{x}, \theta^*, \lambda^* | \theta, \lambda)$



Методи максимальної правдоподібності

- ММП в математичній статистиці – це методи точкового оцінювання невідомих параметрів розподілу шляхом **максимізації функції правдоподібності**. Всі вони базуються на припущенні, що вся інформація про статистичну вибірку знаходиться в цій функції.
- Вони проаналізовані, рекомендовані і значним чином популяризовані **Р. Фішером** у 1912 – 1922 рр., зокрема, у праці «Про математичні основи теоретичної статистики». У ній вперше було використано термін «метод максимальної правдоподібності».



Переваги та недоліки інтегральних методів
максимальної правдоподібності в
порівнянні з іншими методами, що
використовують функції правдоподібності



Переваги:

Інтегрування в порівнянні з максимізацією

$\xi_1, \xi_2, \dots, \xi_n \in N(\theta, 1), \quad \eta \in N(\lambda, \exp(-n\theta^2))$ – незалежні в сукупності ВВ.

Класична ФП:

$$L(\vec{x}, y, \theta, \lambda) = f(x, y | \theta, \lambda) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2\right) (2\pi \exp(-n\theta^2))^{-\frac{1}{2}} \\ \cdot \exp\left(-\frac{(y - \lambda)^2}{2 \exp(-n\theta^2)}\right) = (2\pi)^{-\frac{(n+1)}{2}} \exp\left(-\frac{n}{2} \left(\bar{x}^2 - 2\bar{x}\theta\right) - \frac{(y - \lambda)^2}{2 \exp(-n\theta^2)}\right),$$

Профільна ФП: $\hat{L}(\vec{x}, y, \theta) = f(x, y | \theta, y) = (2\pi)^{-\frac{(n+1)}{2}} \exp\left(-\frac{n}{2} \left(\bar{x}^2 - 2\bar{x}\theta\right)\right) \propto \exp(n\bar{x}\theta)$

Однорідна інтегральна ФП:

$$L^u(\vec{x}, y, \theta) = \int f(\vec{x}, y | \theta, \lambda) d\lambda \propto \exp\left(-\frac{n}{2} (\bar{x} - \theta)^2\right)$$

Профільна ФП швидко прямує до нескінченності, коли $\theta \rightarrow +\infty$

або $\theta \rightarrow -\infty$ (в залежності від знака вибіркового середнього).

Інтегральна ФП має точку глобального максимуму.

Чому можна використовувати знак пропорційності?

Нехай дано $f(x)$, $\alpha = \text{const} > 0$, $\beta = \text{const}$, x_0 – точка максимуму функції $f(x)$

Тоді x_0 – точка максимуму функції $\alpha f(x) + \beta$.

Переваги:

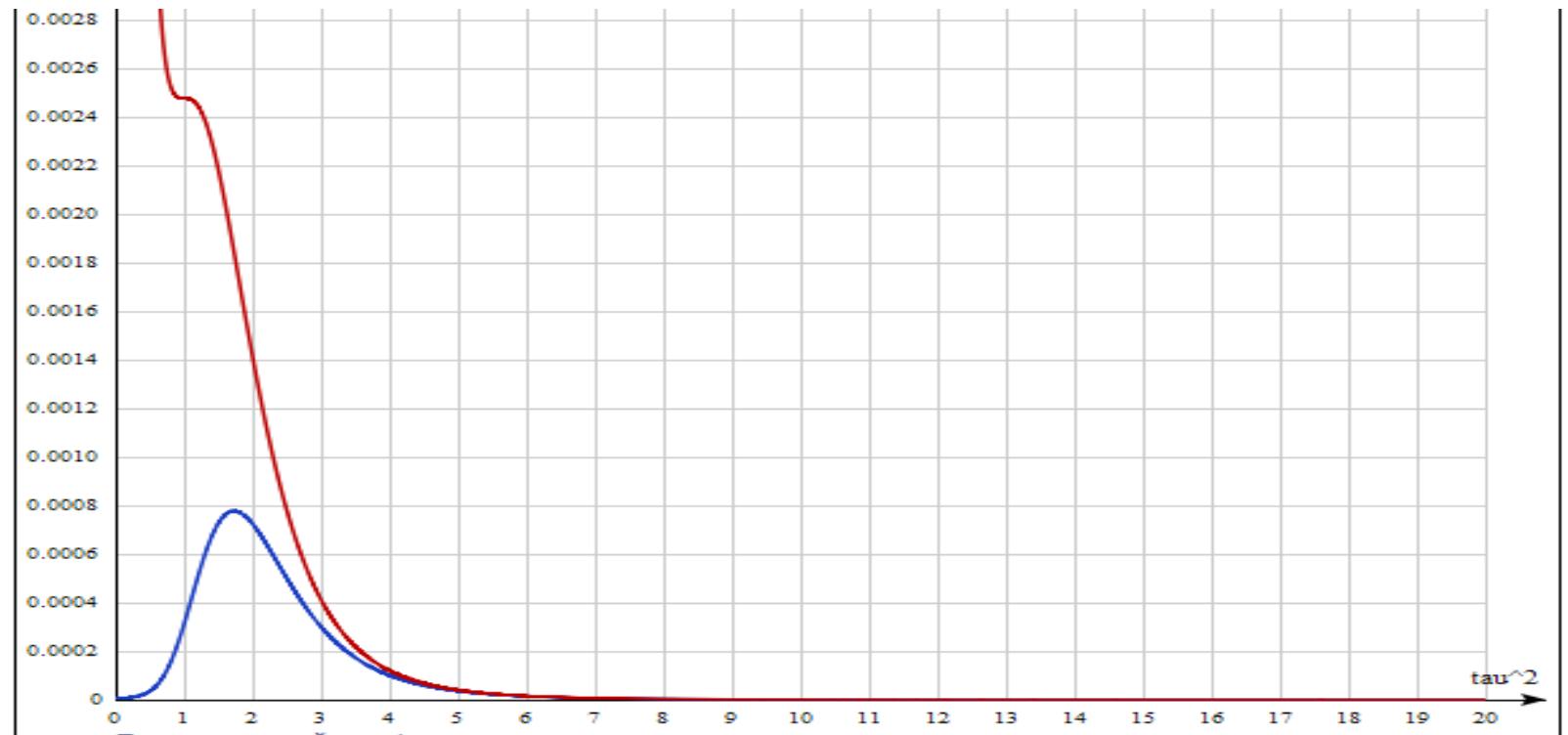
Інтегрування у порівнянні з максимізацією

$\xi_i \in N(\mu_i, 1)$, $\mu_i \in N(\alpha, \tau^2)$. Необхідно оцінити: τ^2 .

В цьому випадку профільна ФП відрізняється від інтегральної тим, що вона має особливу точку в $\tau = 0$ (вона дорівнює $+\infty$ в цій точці). Тому використовуючи профільну ФП, можна зробити хибний висновок, що незалежно від конкретної реалізації вибірки значення точкової оцінки:

$$(\tau^2)^*_{\text{ЗН}} = 0$$

При деяких умовах на вибірку інтегральна та профільна ФП мають наступний вигляд:



Переваги:

Урахування невизначеності несуттєвих параметрів

У профільній ФП несуттєвий параметр замінюється його умовною ОМП.

Інтегральні методи **автоматично враховують невизначеність несуттєвих параметрів** в тому сенсі, що інтегральна ФП – це середня по всіх можливих умовних ФП (за умовою λ).

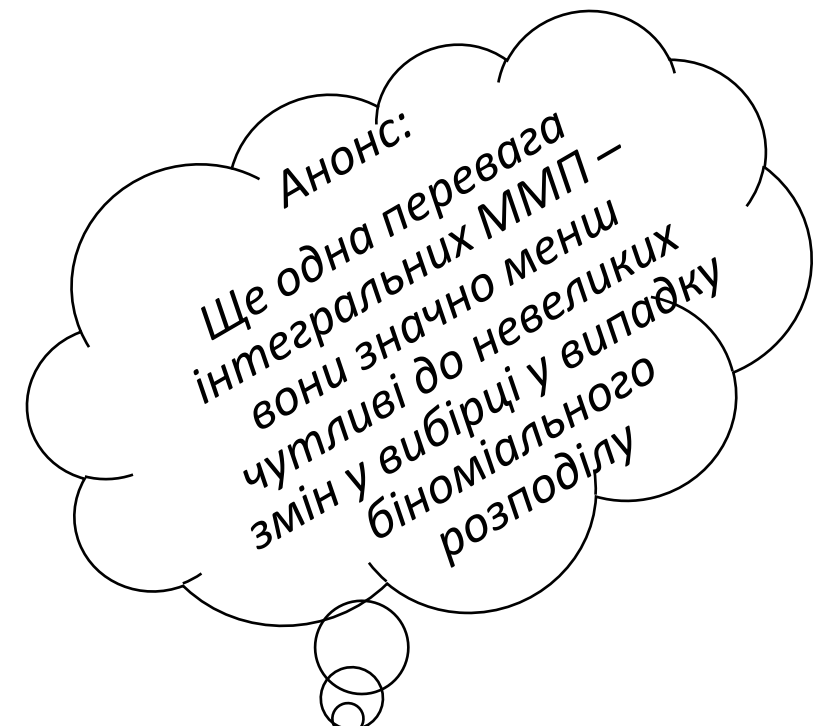
Простота застосування

Після того, як з'явилися комп'ютерні технології, що реалізують **методи Монте-Карло марківських ланцюгів**, чисельне наближення багатовимірних інтегралів перестало бути проблемою.

Проте прості інтегральні методи можуть вимагати трудомісних досліджень для вибору вагової функції.

Аналіз чутливості

Для визначення, наскільки сильно зміниться ФП при зміні дослідних даних, можна **змінювати вагову функцію**.



Недоліки:

Обмеження застосування інтегральних методів

1) **Передчасне усунення несуттєвих параметрів** (втрачена інформація може потім знадобитися при аналізі додаткових даних, що стали доступними пізніше).

2) Проблеми у **метааналізі**:

Нехай $\xi_j^i \in N(\mu_i, \sigma_i^2)$, де $j = \overline{1, n_i}$, $i = \overline{1, p}$ (тобто проведено p стохастичних експериментів і в результаті кожного випробування під номером i отримана конкретна реалізація вибірки обсягом n_i). Нехай $\mu_i \in N(\alpha, \tau^2)$, $i = \overline{1, p}$

Несуттєвий параметр: σ_i^2 . Вагова функція: $\pi(\sigma_i^2) = \frac{1}{\sigma_i^2}$

Інтегральна ФП:

$$L_i(\mu_i) \propto \left(S_i^2 + \left(\bar{X}_i - \mu_i \right)^2 \right)^{-n_i/2}$$

$$\text{де } \bar{X}_i = \sum_{j=1}^{n_i} \frac{X_j^i}{n_i}, \quad S_i^2 = \sum_{j=1}^{n_i} \frac{\left(X_j^i - \bar{X}_i \right)^2}{n_i}, \quad i = \overline{1, p}$$

Підхід: використати добуток ФП + інформацію $\mu_i \in N(\alpha, \tau^2)$, $i = \overline{1, p}$

Але σ_i^2 між собою пов'язані!

Застосування математичних методів на
практиці

Задача оцінювання ефективності роботи
інструмента виявлення шахрайських дій у
платіжній системі



Застосування біноміального закону в оцінці кількості шахрайських дій у платіжній системі

Вибірка – кількість «успіхів» (фактів виявлення ШД) в кожен окремий період.

N – реальна кількість ШД, проведених за період,

p – ймовірність виявити ШД.

Конкретна реалізація вибірки отримана з ГС $\xi \in \text{Bin}(N, p)$.

Отримуємо значення точкової оцінки параметра N ➡

➡ **Визначаємо показник ефективності** роботи програмного забезпечення, призначеного для виявлення ШД:

$$KKД = \frac{X}{N_{3H}^*} \cdot 100\%$$

де X – вибіркове середнє (середнє значення виявлених ШД за період спостереження),

N_{3H}^* – значення точкової оцінки реальної кількості ШД.

Точкове оцінювання параметра N у випадку біноміального розподілу:

Однорідна інтегральна ФП:

$$L^U(\vec{s}, N) = \left(\prod_{j=1}^k C_N^{s_j} \right) \frac{(kN - T)!}{(kN + 1)!}, \quad N \geq s_{\max}$$

Інтегральна ФП Джефрі:

$$L^J(\vec{s}, N) = \left(\prod_{j=1}^k C_N^{s_j} \right) \frac{\Gamma\left(kN - T + \frac{1}{2}\right)}{(kN)!}, \quad N \geq s_{\max}$$

Класична ФП:

$$L(\vec{s}, N, p) = \left(\prod_{j=1}^k C_N^{s_j} \right) p^T q^{Nk-T}, \quad 0 < p < 1, \quad q = 1 - p, \quad N \geq \max_j s_j$$

Профільна ФП:

$$L(\vec{s}, N) = \left(\prod_{j=1}^k C_N^{s_j} \right) \frac{(Nk - T)^{Nk-T}}{(Nk)^{Nk}}, \quad N \geq s_{\max}$$

$$\pi(p) \propto \frac{1}{\sqrt{p(1-p)}}$$

$$\Gamma\left(n + \frac{1}{2}\right) = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2n-3)(2n-1) \frac{\sqrt{\pi}}{2^n}$$

Умовна ФП:

$$L^C(\vec{s}, N) = \frac{\left(\prod_{j=1}^k C_N^{s_j} \right)}{C_{kN}^T}, \quad N \geq s_{\max}$$

Аналіз чутливості

Для аналізу чутливості разом із вихідною вибіркою будемо розглядати ще дві, що відрізняються від неї тільки останнім елементом. У вибірці $\vec{s}_2$ він на 1 більший, а в $\vec{s}_3$ – на 1 менший. Введемо **показник чутливості** для методу:

$$\alpha = \max_i N^*(s_i) - \min_i N^*(s_i)$$

Чим α менший, тим метод є менш чутливим до змін у дослідних даних.

Для вибірки $\vec{s} = (16, 18, 22, 25, 27)$:

Значення точкових оцінок			
	S1	S2	S3
Інтегральна	51	58	45
Джефрі	54	63	46
Профільна	99	192	68
Умовна	infinity	infinity	167

Функція правдоподібності	Показник чутливості α
Однорідна інтегральна ФП	$\alpha = 58 - 45 = 13$
Інтегральна ФП Джефрі	$\alpha = 63 - 46 = 17$
Профільна ФП	$\alpha = 192 - 68 = 124$
Умовна ФП	Неможливо визначити

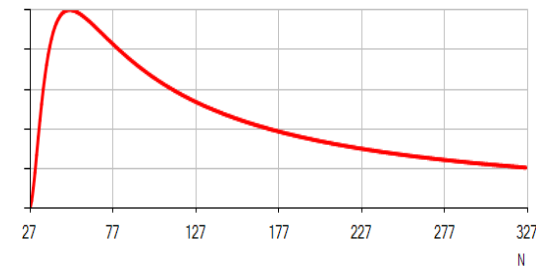
Для реальних даних:

Значення точкових оцінок			
	S1	S2	S3
Інтегральна	18	21	17
Джефрі	19	22	18
Профільна	22	26	20
Умовна	23	27	20

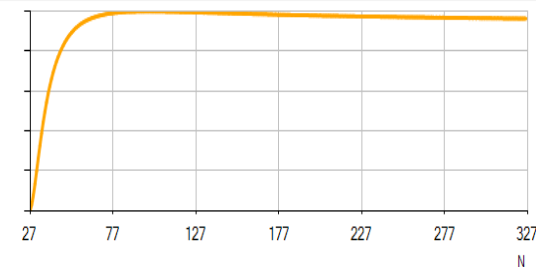
Функція правдоподібності	Показник чутливості α
Однорідна інтегральна ФП	$\alpha = 21 - 17 = 4$
Інтегральна ФП Джефрі	$\alpha = 22 - 18 = 4$
Профільна ФП	$\alpha = 26 - 20 = 6$
Умовна ФП	$\alpha = 27 - 20 = 7$

Вигляд функцій правдоподібності

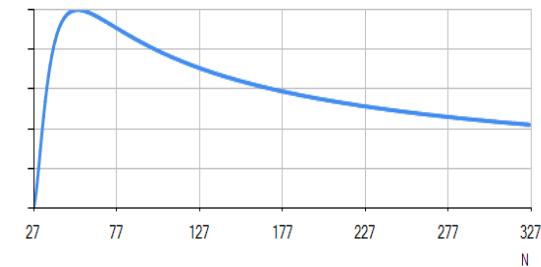
Для невеликої вибірки



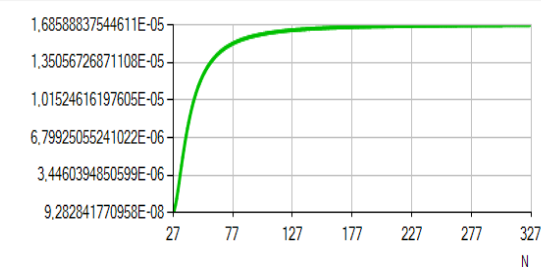
— Однорідна інтегральна ФП



— Профільна ФП

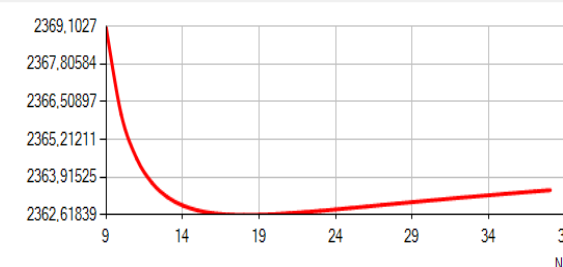


— Інтегральна ФП Джефрі

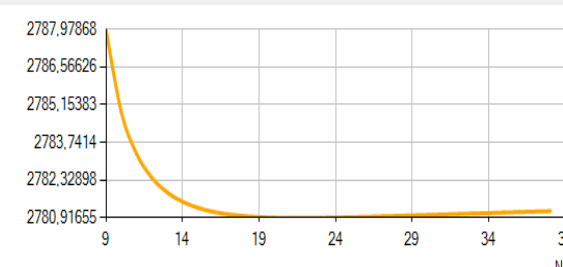


— Умовна ФП

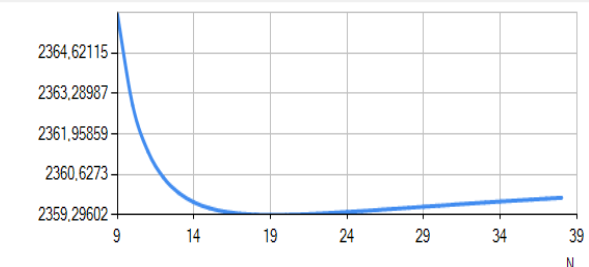
Для реальних даних
(на перших трьох графіках подано
логарифмічні ФП, симетрично відображені
відносно осі абсцис)



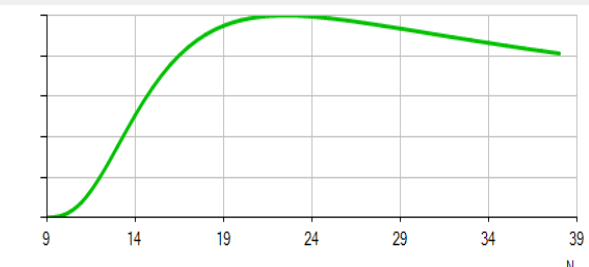
— Однорідна інтегральна ФП



— Профільна ФП



— Інтегральна ФП Джефрі



— Умовна ФП

Відповідь задачі оцінки ефективності інструмента виявлення шахрайства

Коефіцієнт корисної дії програмного забезпечення платіжної системи:

$$KKД = \frac{4,26}{18} \cdot 100\% \approx 23,67\%.$$

Тут 4,26 – значення вибіркового середнього (середній показник, скільки ШД було виявлено програмою), 18 – значення точкової оцінки реальної кількості ШД, отримане однорідним інтегральним ММП.

Отже, є необхідність у подальшому вдосконаленні інструмента або впровадженні нових стратегій та інструментів виявлення ШД.



Висновки

- В дипломній роботі розглянуто інтегральні методи максимальної правдоподібності у контексті інших методів, що використовують функції правдоподібності. Вивчено підходи представників різних наукових шкіл до означення функції правдоподібності та до різноманітних способів її застосування.
- Отримано точкові оцінки параметра N біноміального розподілу різними методами максимальної правдоподібності і встановлено, що інтегральні оцінки **менш чутливі до незначних змін у вибірках**, ніж профільні та умовні.
- Створено **програмний продукт**, який здійснює пошук значень точкових оцінок та ілюструє на графіках проблеми, які можуть виникати при точковому оцінюванні в дискретному випадку (зокрема, зростаюча умовна функція правдоподібності та поведінка профільної ФП як константи на великому проміжку).
- Визначено **переваги** інтегральних методів, які дозволяють використовувати їх на практиці, а також **обмеження у застосуванні**, зумовлені здебільшого тим, що в несуттєвих параметрах міститься багато інформації про параметри, що оцінюються.

Перспективи для подальших досліджень

- Детальний розгляд байєсівського підходу до точкового оцінювання методами максимальної правдоподібності.
- Дослідження питання вибору доцільної вагової функції для інтегральних методів максимальної правдоподібності.
- Отримання точкових оцінок в умовах наявності несуттєвих параметрів для інших законів розподілу: гамма-розподілу, бета-розподілу, розподілу Вейбулла та ін.

Дякую за увагу!

