Доронін Олексій В'ячеславович. Назва дисертаційної роботи: "Семіпараметричне оцінювання та перевірка гіпотез за спостереженнями у суміші"

Київський національний університет імені Тараса Шевченка

На правах рукопису

ДОРОНІН ОЛЕКСІЙ В’ЯЧЕСЛАВОВИЧ

УДК 519.21

СЕМІПАРАМЕТРИЧНЕ ОЦІНЮВАННЯ

ТА ПЕРЕВІРКА ГІПОТЕЗ

ЗА СПОСТЕРЕЖЕННЯМИ У СУМІШІ

01.01.05 — теорія ймовірностей і математична статистика

Дисертація на здобуття наукового ступеня

кандидата фізико-математичних наук

Науковий керівник

Майборода Ростислав Євгенович,

доктор фізико-математичних наук, професор

Київ — 2015

1

ЗМІСТ

Вступ 5

Актуальність теми . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

Зв’язок роботи з науковими програмами, планами, темами . . . . . . 6

Мета і завдання дослідження . . . . . . . . . . . . . . . . . . . . . . . 6

Методи дослідження . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

Наукова новизна одержаних результатів . . . . . . . . . . . . . . . . 7

Практичне значення одержаних результатів . . . . . . . . . . . . . . . 8

Особистий внесок здобувача . . . . . . . . . . . . . . . . . . . . . . . 8

Апробація результатів дисертації . . . . . . . . . . . . . . . . . . . . . 8

Публікації . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

Зміст роботи . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

Основні позначення та домовленості . . . . . . . . . . . . . . . 11

Попередні результати . . . . . . . . . . . . . . . . . . . . . . . . 13

Непараметричне оцінювання . . . . . . . . . . . . . . . . . . . . 14

Семіпараметричне оцінювання . . . . . . . . . . . . . . . . . . . 17

Параметричне оцінювання . . . . . . . . . . . . . . . . . . . . . 23

Тести для перевірки гіпотез . . . . . . . . . . . . . . . . . . . . . 26

Імітаційне моделювання . . . . . . . . . . . . . . . . . . . . . . 29

Розділ 1. Огляд літератури 33

1.1. Модель суміші зі змінними концентраціями . . . . . . . . . . . 33

1.1.1. Історія розвитку моделі суміші . . . . . . . . . . . . . . 33

1.1.2. Сучасні дослідження моделі суміші зі змінними концентраціями . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

1.2. Попередні результати . . . . . . . . . . . . . . . . . . . . . . . . 36

1.2.1. Навантажені емпіричні міри . . . . . . . . . . . . . . . . 36

2

1.2.2. Мінімаксні вагові коефіцієнти . . . . . . . . . . . . . . . 38

1.2.3. Виправлені вагові коефіцієнти . . . . . . . . . . . . . . . 40

1.2.4. Оцінювання квантилів . . . . . . . . . . . . . . . . . . . 43

1.2.5. Оцінювання моментів . . . . . . . . . . . . . . . . . . . . 45

1.3. Допоміжні відомості . . . . . . . . . . . . . . . . . . . . . . . . . 46

1.3.1. Умовна дисперсія та коваріація . . . . . . . . . . . . . . 46

1.3.2. Збіжність перетворень . . . . . . . . . . . . . . . . . . . 47

1.3.3. Закон великих чисел . . . . . . . . . . . . . . . . . . . . 48

1.3.4. Центральна гранична теорема . . . . . . . . . . . . . . . 48

1.3.5. Регулярні статистичні експерименти. Нерівність КрамераРао . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48

Розділ 2. Непараметричне оцінювання 52

2.1. Оцінювання моментів . . . . . . . . . . . . . . . . . . . . . . . . 52

2.1.1. Асимптотична нормальність невиправлених оцінок моментів . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52

2.1.2. Асимптотика виправлених оцінок моментів . . . . . . . 57

2.1.3. Оцінка знизу для коваріаційної матриці оцінок моментів 62

2.2. Оцінювання розподілів . . . . . . . . . . . . . . . . . . . . . . . 64

2.2.1. Асимптотика невиправлених оцінок розподілів . . . . . 65

2.2.2. Асимптотика виправлених оцінок розподілів . . . . . . . 65

2.3. Оцінювання квантилів . . . . . . . . . . . . . . . . . . . . . . . . 67

Висновки до другого розділу . . . . . . . . . . . . . . . . . . . . . . . 70

Розділ 3. Семіпараметричне оцінювання 71

3.1. Оцінки методом моментів . . . . . . . . . . . . . . . . . . . . . . 72

3.2. GEE-оцінки . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78

3.2.1. Визначення . . . . . . . . . . . . . . . . . . . . . . . . . 78

3.2.2. Консистентність . . . . . . . . . . . . . . . . . . . . . . . 81

3.2.3. Асимптотична нормальність . . . . . . . . . . . . . . . . 83

3.3. Нижня межа матриці розсіяння GEE-оцінок . . . . . . . . . . . 87

3

3.3.1. Обмеження без втрати загальності . . . . . . . . . . . . . 87

3.3.2. Асимптотична поведінка незсунутих нормованих оцінок 91

3.3.3. Вагові коефіцієнти у загальному вигляді . . . . . . . . . 93

3.3.4. Незсунуті вагові коефіцієнти . . . . . . . . . . . . . . . . 98

3.4. Адаптивні оцінки . . . . . . . . . . . . . . . . . . . . . . . . . . 101

3.4.1. Визначення . . . . . . . . . . . . . . . . . . . . . . . . . 102

3.4.2. Асимптотична поведінка . . . . . . . . . . . . . . . . . . 104

3.5. Оптимальні адаптивні оцінки . . . . . . . . . . . . . . . . . . . . 106

3.5.1. Мінімізація матриці розсіяння . . . . . . . . . . . . . . . 106

3.5.2. Оцінювання матриці розсіяння . . . . . . . . . . . . . . . 111

3.5.3. Реалізація оптимальної оцінки . . . . . . . . . . . . . . . 114

Висновки до третього розділу . . . . . . . . . . . . . . . . . . . . . . 116

Розділ 4. Параметричне оцінювання 118

4.1. Позначення . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118

4.2. Регулярні експерименти у моделі суміші . . . . . . . . . . . . . 120

4.3. Оцінка максимальної вірогідності . . . . . . . . . . . . . . . . . 122

Висновки до четвертого розділу . . . . . . . . . . . . . . . . . . . . . 125

Розділ 5. Тести для перевірки гіпотез 127

5.1. Постановка задачі . . . . . . . . . . . . . . . . . . . . . . . . . . 127

5.1.1. Гіпотези про функціональні моменти компонентів суміші 128

5.1.2. Гіпотеза про рівність характеристик розподілів компонентів . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 129

5.2. Побудова тесту . . . . . . . . . . . . . . . . . . . . . . . . . . . . 130

5.2.1. Асимптотика тестової функції . . . . . . . . . . . . . . . 130

5.2.2. Визначення тесту . . . . . . . . . . . . . . . . . . . . . . 132

5.3. Перевірка гіпотези про рівність розподілів . . . . . . . . . . . . 132

5.3.1. Постановка задачі . . . . . . . . . . . . . . . . . . . . . . 132

5.3.2. Дослідження тестової статистики . . . . . . . . . . . . . 134

Висновки до п’ятого розділу розділу . . . . . . . . . . . . . . . . . . . 136

4

Розділ 6. Імітаційне моделювання 138

6.1. Оцінювання параметрів . . . . . . . . . . . . . . . . . . . . . . . 138

6.1.1. Постановка моделювання . . . . . . . . . . . . . . . . . 138

6.1.2. Деталі реалізації оцінок . . . . . . . . . . . . . . . . . . . 142

6.1.3. Результати моделювання . . . . . . . . . . . . . . . . . . 146

6.1.4. Загальні висновки . . . . . . . . . . . . . . . . . . . . . . 158

6.2. Перевірка гіпотез . . . . . . . . . . . . . . . . . . . . . . . . . . 159

6.2.1. Гіпотези про функціональні моменти . . . . . . . . . . . 160

6.2.2. Тест хі-квадрат . . . . . . . . . . . . . . . . . . . . . . . 164

6.2.3. Загальні висновки . . . . . . . . . . . . . . . . . . . . . . 168

6.3. Аналіз реальних даних . . . . . . . . . . . . . . . . . . . . . . . 168

6.3.1. Опис даних . . . . . . . . . . . . . . . . . . . . . . . . . 169

6.3.2. Параметризація компонентів . . . . . . . . . . . . . . . . 171

6.3.3. Застосування тестів . . . . . . . . . . . . . . . . . . . . . 172

6.3.4. Оцінювання параметрів . . . . . . . . . . . . . . . . . . . 176

6.3.5. Загальні висновки . . . . . . . . . . . . . . . . . . . . . . 179

Висновки до шостого розділу розділу . . . . . . . . . . . . . . . . . . 180

Висновки 181

Список використаних джерел 183

5

ВСТУП

Актуальність теми

Модель суміші ймовірнісних розподілів виникає у випадку, коли дослідник

може віднести певний спостережуваний об’єкт до деякої підпопуляції лише з

певною долею впевненості (така проблема називається проблемою неповної інформації). При цьому кожна окрема підпопуляція відображає один компонент

суміші, а ймовірність того, що об’єкт належить цій популяції, називається концентрацією компонента. Моделі сумішей, у яких концентрації є сталими для

всієї вибірки, називаються класичними моделями сумішей. Вони виникли ще

у XIX столітті у працях Ньюкомба [58] та Пірсона [60]. Протягом XX століття дана модель досліджувалась у багатьох роботах. Огляди таких досліджень

можна знайти у монографіях Маклахлана та Басфорда [55], Маклахлана та Піла [56]. Сучасні дослідження у цій галузі представлені роботами Ньгуйєна та

Маклахлана [59], Дугласа [61], і збірнику Регаззоні та Джонса [62].

Природним узагальненням класичної моделі суміші є модель суміші зі змінними концентраціями, де концентрації компонентів змінюються в залежності

від спостереження. У монографіях Р. Майбороди та О. Сугакової [17, 18] зібрано останні результати з теорії моделі суміші зі змінними концентраціями. Дослідження зарубіжних авторів із даної тематики представлено роботами Халла

і Тіттерінгтона [43], Тіттерінгтона, Сміта і Макова [71], і т.д.

Дисертаційна робота присвячена оцінюванню параметрів розподілів компонентів у семіпараметричній моделі суміші зі змінними концентраціями (коли

параметричні сімейства розподілів задані лише для деяких компонентів суміші), а також перевірці гіпотез стосовно розподілів компонентів або їх числових

характеристик (рівність середніх, дисперсій, тощо). Результати даної роботи

можуть знайти застосування при аналізі даних широкого кола досліджень (ме-

6

дичних, біологічних, соціологічних, психологічних, політологічних, економічних, тощо).

Зв’язок роботи з науковими програмами, планами, темами

Дисертаційна робота виконана в рамках державної бюджетної науководослідницької теми № 11БФ038-02 “Еволюційні системи: дослідження аналітичних перетворень, випадкових флуктуацій та статистичних закономірностей”

(номер державної реєстрації 0111U006561) кафедри теорії ймовірностей, статистики та актуарної математики механіко-математичного факультету Київського національного університету імені Тараса Шевченка, що входить до комплексного тематичного плану науково-дослідних робіт “Сучасні математичні проблеми природознавства, економіки та фінансів”.

Мета і завдання дослідження

Мета роботи – створення нових, більш ефективних алгоритмів оцінювання

та перевірки гіпонез у непараметричних та семіпараметричних моделях сумішей зі змінними концентраціями, а також аналіз якості отриманих алгоритмів.

Основними завданнями дослідження є:

▷ Побудувати робастну оцінку для параметру розподілів компонентів суміші зі змінними концентраціями та дослідити її асимптотичну поведінку.

▷ Побудувати оцінки методу узагальнених оціночних рівнянь (GEE) у моделі суміші зі змінними концентраціями, а також адаптивні оцінки на їх

основі; дослідити асимптотичну поведінку отриманих оцінок.

▷ Знайти нижню межу матриць розсіяння побудованих оцінок. Визначити

оцінки, на яких вона досягається.

▷ Порівняти знайдену нижню межу із інформаційною межею Фішера.

▷ Розробити тести для перевірки гіпотез про функціональні моменти розподілів компонентів суміші. Дослідити їх асимптотичну поведінку.

7

▷ Розробити тести для перевірки рівності розподілів компонентів по групованим даним.

▷ Дослідити роботу отриманих оцінок і тестів на модельованих даних.

Методи дослідження

У дисертації використані методи теорії ймовірностей, математичної статистики, математичного аналізу, теорії міри, функціонального аналізу.

Наукова новизна одержаних результатів

▷ Доведено асимптотичну нормальність сукупного вектора квантилів для

різних компонентів суміші.

▷ Побудована оцінка методом моментів у випадку, коли евклідовим параметром параметризовані одночасно декілька розподілів компонентів

суміші (багатокомпонентна семіпараметрична модель).

▷ Визначені оцінки методу узагальнених оціночних рівнянь (GEE-оцінки)

у випадку багатокомпонентної семіпараметричної моделі суміші.

▷ Розглянуто адаптивні оцінки, які наближають GEE-оцінки за допомогою розкладу останніх за певним функціональним базисом.

▷ Визначено оцінку максимальної вірогідності у випадку, коли розподіли

всіх компонентів суміші задані параметричним сімейством. Знайдено

інформаційну межу за умов регулярності моделі.

▷ Для всіх введених оцінок доведено консистентність та асимптотичну

нормальність.

▷ Знайдено нижню межу матриці розсіяння GEE-оцінок за умов незсуненості GEE-рівняння. Визначено адаптивні оцінки, які наближають вказану нижню межу.

▷ Досліджено якість отриманих оцінок на модельованих даних.

▷ Розроблено тести для перевірки гіпотез про функціональні моменти

розподілів компонентів. Досліджена їх асимптотична поведінка.

8

▷ Для перевірки рівності розподілів кількох компонентів розроблений

аналог класичного тесту χ

2

.

▷ Робота тестів перевірена на модельованих даних.

Практичне значення одержаних результатів

Всі результати дисертаційної роботи є теоретичними. Практичне значення

вони набувають при аналізі даних різного роду досліджень (медичних, біологічних, соціологічних, психологічних, політологічних, економічних, тощо).

Особистий внесок здобувача

За результатами дисертації було опубліковано шість статей у фахових журналах. У роботі [54] автору належать розрахунки асимптотичних коефіцієнтів

росіяння та проведення імітаційного моделювання, результати інших п’яти робіт були отримані автором самостійно. Результати [54] були узагальнені автором у роботах [4, 5]. Робота [41] була опублікована у співавторстві з науковим

керівником Майбородою Р.Є., якому належить постановка задачі та загальне

керівництво роботою. Чотири роботи ( [3–5, 7]) опубліковані здобувачем без

співавторів.

Апробація результатів дисертації

Результати дисертаційної роботи обговорювались на наступних зустрічах.

▷ Всеукраїнський конкурс студентських наукових робіт з математичних

наук (Івано-Франківськ, Україна, 2012).

▷ Третя міжнародна конференція “Modern Stochastics: Theory and Applications”

(Київ, Україна, 2012).

▷ Вісімнадцята міжнародна конференція молодих статистиків “European

Young Statisticians Meeting” (Osijek, Croatia, 2013).

9

▷ Десята міжнародна конференція “Computer Data Analysis & Modeling”

(Мінськ, Білорусь, 2013).

▷ П’ятнадцята міжнародна наукова конференція імені академіка Михайла

Кравчука (Київ, Україна, 2014).

▷ Міжнародна конференція “Multivariate Statistical Analysis” (Lodz, Poland,

2014).

▷ Міжнародна конференція “Probability, reliability and stochastic optimization”

(Київ, Україна, 2015).

▷ Засідання наукового семінару “Асимптотичні методи статистики” на

механіко-математичному факультеті КНУ імені Тараса Шевченка під

керівництвом професорів О. Г. Кукуша та Р. Є. Майбороди (Київ, Україна, 2012–2015).

▷ Засідання наукового семінару “Статистичні проблеми для випадкових

процесів і полів” при кафедрі математичного аналізу та теорії ймовірностей НТУУ “КПІ” (Київ, Україна, 2015).

▷ Засіданні наукового семінару “Теорія стохастичних процесів та її застосування” при кафедрі теоретичної та прикладної статистики ЛНУ

(Львів, Україна, 2015).

▷ Засіданні наукового семінару кафедри теорії ймовірностей, статистики

та актуарної математики КНУ імені Тараса Шевченка (Київ, 2015).

Публікації

За результатами дисертації здобувач опублікував шість робіт у фахових виданнях [3–5, 7, 41, 54] і тези шести доповідей на конференціях [6, 38–40, 42, 53].

Зміст роботи

Робота складається зі вступу, шести розділів, розбитих на підрозділи, висновків та списку літератури.

10

У вступі розкривається актуальність теми дисертації, мета, завдання та

методи дослідження. Обговорюється наукова новизна та практичне значення

отриманих результатів. Вказується особистий внесок здобувача у дослідження, згадуються публікації, розкривається зміст роботи.

Перший розділ присвячено огляду літератури. Коротко наводиться історія

розвитку моделі суміші зі змінними концентраціями. Вказуються досягнення за

темою дисертації, які вже були отримані в інших роботах. Розділ закінчується

додатковими загальновідомими результатами теорії ймовірностей і математичної статистики, які використовуються у даній роботі.

Другий розділ присвячено оцінкам функціональних моментів і квантилів

у моделі суміші. Авторським досягненням є узагальнення на випадок вектора

емпіричних моментів або квантилів, до якого входять елементи з різних компонентів суміші. Отримані результати використовуються у наступних розділах.

У третьому розділі обговорюється задача семіпараметричного оцінювання.

Припускається, що для частини компонентів задане деяке сімейство їх розподілів, що описується евклідовим параметром. Розподіли інших компонентів

вважаються повністю невідомими. Мета – оцінити даний параметр. Для цього розглядаються три види оцінок: методом моментів, узагальнених оціночних

рівнянь (GEE), і адаптивних. Для всіх оцінок доводиться консистентність та

асимптотична нормальність; для GEE- та адаптивних оцінок знаходиться нижня межа їх матриці розсіяння, та оцінки, на яких вона досягається.

Четвертий розділ присвячено випадку, коли розподіли всіх компонентів можна описати за допомогою одного евклідового параметра. Знаходиться оцінка

максимальної вірогідності для моделі суміші зі змінними концентраціями, а також відповідна інформаційна матриця за умов регулярного експерименту.

У п’ятому розділі йдеться про задачу перевірки гіпотез стосовно функціональних моментів розподілів суміші зі змінними концентраціями. У рамках даної задачі визначається тестова статистика, досліджується її асимптотична поведінка. Будується тест для перевірки гіпотез. Реалізується тест для перевірки

11

рівності розподілів компонентів по групованим даним, який є аналогом класичного тесту хі-квадрат.

У шостому розділі поміщено опис і результати імітаційного моделювання,

метою якого була перевірка роботи розроблених підходів. Проводиться аналіз

реальних даних (результатів парламентських виборів) за допомогою розроблених технік.

Далі і до кінця даного розділу вміщено огляд основних результатів дисертаційного дослідження

ВИСНОВКИ

ДисертаційнароботаприсвяченадослідженнюмоделісумішіКонцентрації

компонентівсумішівважаютьсявідомимиіможутьзмінюватисявідспостереженнядоспостереженняДляданоїмоделірозглянутозадачіоцінюванняевклідовогопараметрусімействарозподілівкомпонентівіперевіркигіпотезпро

функціональнімоментикомпонентівсуміші

ЗначнаувагаприділяєтьсязадачісеміпараметричногооцінюванняВостаннійвважаєтьсящодлядеякихкомпонентівсумішізаданепараметричнесімействорозподілівякеїхописуєрозподілирештикомпонентіввважаютьсяповністюневідомимиПобудованіоцінкиметодоммоментівіузагальненихоціночнихрівняньЗапропонованіадаптивні

оцінкиметоюякихємаксимальненаближенняоцінокДлявведенихоцінокдоведеноїхконсистентністьтаасимптотичнунормальністьЗнайденонижнюмежуїхматрицірозсіянняіоцінкинаякихнижнямежадосягаєтьсяДля

адаптивнихоцінокзапропонованоалгоритмїхреалізації

ДодатковорозглянутовипадокколирозподіливсіхкомпонентівописуютьсязадопомогоюевклідовогопараметруЗнайденонижнюмежуматрицірозсіянняасимптотичнонезсунутихоцінокзаумоврегулярностіЗапропоновано

лінеаризовануоцінкумаксимальноївірогідностінаякійдосягаєтьсязнайдена

асимптотичнамежа

Длязадачіперевіркигіпотезпрофункціональнімоментизапропоновано

тестдосягнутийрівеньзначущостіякогоасимптотичнопрямуєдотеоретичногозначенняЯкчастковийвипадокрозглянутогіпотезупрорівністьпари

розподілівнагрупованихданихОстанняперевіряєтьсязадопомогоютесту

аналогічногодокласичноготестухіквадрат

Роботузапропонованихоцінокітестівбулоперевіренонамодельованих

данихРозглянутоаналізреальнихданихнаприкладірезультатівукраїнських



парламентськихвиборіврокуЗапропонованітехнікиможутьбутизастосованіуширокомуколінауковихдослідженьмедичнібіологічнісоціологічніпсихологічніполітологічніекономічнітощо