МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

МАТЕМАТИЧНИЙ ФАКУЛЬТЕТ
Кафедра фундаментальної та прикладної математики

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему: **«АНАЛІЗ ДАНИХ З ВИКОРИСТАННЯМ БУТСТРЕП-ПРОЦЕДУРИ»**

Виконав(ла): студент(ка) 2 курсу, групи 8.1130-з

спеціальності 113 прикладна математика
(шифр і назва спеціальності)

освітньої програми прикладна математика
(назва освітньої програми)

К.О. Драганова
(ініціали та прізвище)

Керівник доцент кафедри фундаментальної та прикладної математики, доцент к.ф.-м.н. Швидка С.П.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Рецензент доцент кафедри програмної інженерії, доцент к.ф.-м.н. Кудін О.В.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Запоріжжя – 2022

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет математичний

Кафедра фундаментальної та прикладної математики

Рівень вищої освіти магістр

Спеціальність 113 прикладна математика

(шифр і назва)

Освітня програма прикладна математика

ЗАТВЕРДЖУЮ

Завідувач кафедри фундаментальної та прикладної математики, д.т.н., доцент

_____ Гребенюк С.Н.

(підпис)

« ____ » _____ 2021 р.

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТОВІ(СТУДЕНТЦІ)

Драгановій Катерині Олександрівні

(прізвище, ім'я та по-батькові)

1. Тема роботи (проекту) Аналіз даних з використанням бутстреп-процедури

керівник роботи (проекту) Швидка Світлана Петрівна, к.ф.-м.н., доцент

(прізвище, ім'я та по-батькові, науковий ступінь, вчене звання)

затверджені наказом ЗНУ від «____» _____ 2021 року № _____

2. Строк подання студентом роботи _____

3. Вихідні дані до роботи 1. Постановка задачі
2. Перелік літератури

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

Теоретичні відомості про бутстреп та його виникнення.

Аналіз даних з використанням бутстреп-процедури.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) _____

Презентація, графіки

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	Доцент Швидка С.П.		
2	Доцент Швидка С.П.		

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1.	Розробка плану роботи.		
2.	Збір вихідних даних.		
3.	Обробка методичних та теоретичних		
	джерел.		
4.	Розробка першого розділу.		
5.	Розробка другого розділу.		
6.	Оформлення та нормоконтроль		
	кваліфікаційної роботи.		
7.	Захист кваліфікаційної роботи.		

Студент _____ К. О. Драганова
(підпис) (ініціали та прізвище)

Керівник роботи _____ С.П. Швидка
(підпис) (ініціали та прізвище)

Нормоконтроль пройдено

Нормоконтролер _____ В.В. Леонтьєва
(підпис) (ініціали та прізвище)

РЕФЕРАТ

Кваліфікаційна робота магістра «Аналіз даних з використанням бутстреп-процедури»: 45 с., 9 рисунків, 1 таблиця, 16 джерел.

R, BOOTSTRAP, STATISTICS, DATA ANALYSIS.

Об’єкт дослідження – статистичні дані.

Мета роботи – дослідження можливостей бутстреп-процедури для аналізу даних.

Методи дослідження: статистичний аналіз, моделювання.

Предмет дослідження: побудова довірчих інтервалів для вибіркового середнього за допомогою бутстреп-процедури.

Бутстреп – це статистична процедура, яка повторює вибірку одного набору даних для створення багатьох змодельованих зразків. Цей процес дозволяє обчислювати стандартні помилки, будувати довірчі інтервали та виконувати перевірку гіпотез для численних типів вибіркової статистики. Методи бутстреп є альтернативними підходами до традиційної перевірки гіпотез і відрізняються тим, що їх легше зрозуміти і дійсні для більшої кількості умов.

Цей метод має рівну ймовірність випадкового відображення кожної вихідної точки даних для включення в набори даних із повторною вибіркою. Процедура може вибрати точку даних більше одного разу для набора даних із повторною вибіркою. Ця властивість є аспектом процесу «із заміною». Процедура створює повторно вибіркові набори даних, які мають той самий розмір, що й вихідний набір даних.

Процес закінчується тим, що змодельовані набори даних мають багато різних комбінацій значень, які існують у вихідному наборі даних.

SUMMARY

Master's qualifying paper "Data Analysis by Using Bootstrap Procedure": 45 pages, 9 figures, 1 table, 16 references.

R, BOOTSTRAP, STATISTICS, DATA ANALYSIS

The object of research is a Zaporozhzhia air pollution data.

The purpose of the work is to study the possibilities of the bootstrap procedure for data analysis.

Research methods: analysis, modeling.

Subject of research: construction of confidence intervals on the basis of samples created with the help of bootstrap.

Bootstrap is a statistical procedure that replicates the sampling of one set of data to create many simulated samples. This process allows you to calculate standard errors, build confidence intervals, and test hypotheses for many types of sample statistics. Bootstrap methods are alternative approaches to traditional hypothesis testing and differ in that they are easier to understand and valid for more conditions.

This method has an equal probability of randomly displaying each data source for inclusion in re-sampled datasets. The procedure can select a data point more than once for a re-sampled data set. This property is an aspect of the "replacement" process. The procedure recreates sample datasets that are the same size as the original dataset.

The process ends with the simulated datasets having many different combinations of values that exist in the original dataset.

ЗМІСТ

ВСТУП

Прикладна статистика бурхливо розвивається останні десятиліття. Серйозним стимулом є продуктивність обчислювальних засобів, що стрімко зростає. Тому зрозумілий гострий інтерес до статистичних методів із використанням комп'ютерів.

Від традиційних методів бутстреп відрізняється тим, що він передбачає багаторазову обробку різних частин тих самих даних. Підхід був запропонований Б.Ефроном в 1977 р.

Сам термін «бутстреп» (bootstrap) буквально означає «витягування себе за шнурки від черевиків», метод бутстреп – подальший розвиток «методу складного ножа». Ідея, яку запропонував у 1949 році М. Кенуй («метод складаного ножа») полягає у тому, щоб з однієї вибірки зробити багато шляхом послідовного виключення одного спостереження (і повертаючи його у вибірку перед вибором іншої підвибірки). Б. Ефрон розробив новий спосіб розмноження вибірок, що істотно використовує датчики псевдовипадкових чисел. Він запропонував будувати нові вибірки, моделюючи вибірки із емпіричного розподілу.

Бутстреп з'явився для того, щоб боротися зі зміщенням, зумовленим вибіркою. Бутстреп придатний до роботи з будь-якими статистичними завданнями, чи це перевірки гіпотез про закони розподілу випадкових величин, регресія, дисперсійний аналіз чи багатовимірна класифікація даних.

Але там, де раніше виникали різні труднощі, тепер нам на допомогу приходить бутстреп. Довгий час не вдавалося подолати статистичні труднощі у моделях нелінійної за параметрами регресії. У даний час побудова довірчих інтервалів є найважливішою практичною стороною використання цього методу.

Як відомо, метою статистичних методів служить представлення отриманих даних у компактному, зручному та наочному вигляді, узагальнення

їх за допомогою математичних моделей та прийняття рішень про оптимальні подальші дії.

У нових науково-практичних задачах зі складними алгоритмами, властивості яких недостатньо зрозумілі, бутстреп є цінним інструментом для вивчення ситуації.

Бутстреп відрізняється від традиційних методів тим, що він передбачає багаторазову обробку різних частин тих самих даних, як би поворот їх «різними гранями», і зіставлення отриманих таким чином результатів.

Вищенаведене обумовлює актуальність теми кваліфікаційної роботи, яка полягає у аналізі даних шляхом бутстреп-процедури.

Дипломна робота складається з двох частин та додатку.

Перша частина включає характеристику методів рандомізованої обробки даних.

Друга частина включає застосування бутстреп методу до аналізу емпіричних даних.

Додаток включає програмний код на мові R.

1 МЕТОДИ РАНДОМІЗОВАНОЇ ОБРОБКИ ДАНИХ

1.1 Виникнення бутстрепу

Бутстреп-процедура (або bootstrap) була запропонована Б. Ефроном [7], як деяке узагальнення алгоритму «складного ножа», «щоб не зменшувати щоразу кількість елементів у порівнянні з вихідною сукупністю». За однією з версій слово bootstrap означає шкіряну смужку у вигляді петлі, що прикріплюється до задника похідного черевика для полегшення його натягування на ногу.

Основна ідея бутстрепу по Б. Ефрону [7] полягає у тому, щоб методом статистичних випробувань Монте-Карло багаторазово витягувати повторні вибірки з емпіричного розподілу. А саме, береться кінцева сукупність з n членів вихідної вибірки $x_1, x_2, ..., x_{n-1}, x_n$, з якої на кожному кроці з n послідовних ітерацій за допомогою датчика випадкових чисел рівномірно розподілених на інтервалі [1, n] «витягується» довільний елемент, який знову «повертається» у вихідну вибірку (тобто може бути витягнутий повторно). Наприклад, при n = 6 одна з таких комбінацій має вигляд x_4, x_2, x_2, x_4, x_5, тобто одні елементи можуть повторюватися два або більше разів, тоді як інші елементи відсутні. У такий спосіб можна сформувати будь-яку, скільки завгодно велику кількість бутстреп-вибірок (зазвичай 5000 – 10000). Як і у випадку «складного ножа», у результаті легкої модифікації частотного розподілу реалізацій вихідних даних можна очікувати, що кожна наступна псевдовибірка, що генерується, буде повертати значення параметра, які трохи відрізняються від обчисленого параметру для початкової сукупності. На основі розкиду значень аналізованого показника, отриманого у процесі імітації, можна побудувати, наприклад, довірчі інтервали параметра, що оцінюється. Тим самим «бутстреп є більш економним способом статистичного дослідження, що використовує

переваги застосування комп'ютера та дозволяє обійтися без додаткових натурних вимірювань».

Бутстрапування – це будь-який тест чи показник, який використовує випадкову вибірку із заміною (наприклад, імітує процес вибірки) і підпадає під ширший клас методів повторної вибірки. Бутстрапування надає можливість оцінити рівень точності (систематична помилка, дисперсія, довірчі інтервали, помилка прогнозування тощо) для вибіркових оцінок. Цей метод дозволяє оцінити вибіркові характеристики практично будь-якої статистики з використанням методів випадкової вибірки [8 – 16].

Алгоритм, запропонований М. Кенуем у 1949 р., полягав у тому, щоб послідовно вилучати з наявної вибірки даних спостереження або значення вимірюваної величини по одному спостереженню чи значенню. Наступний крок полягав у тому, щоб провести обробку всієї інформації, що залишилася, і передбачити результат у виключеній точці. Сукупність розбіжностей, отриманих таким чином по всіх точках, несе у собі інформацію про вибіркове зміщення, якою можна скористатися для уточнення характеристик. Дж. Тьюкі активно удосконалив цей метод і дав йому назву «jackknife» (складний ніж), і використав для оцінки дисперсії, математичного сподівання сукупності вибірок, що вивчається, і перевірки нульової гіпотези про те, що розподіл деякої статистичної величини є симетричним щодо заданої точки. Поняття «складний ніж» відноситься до універсального методу, який покликаний замінити особливі методики, які не завжди придатні на практиці, подібно до бойскаутського ножа, що годиться на всі випадки життя.

Б. Ефрон описав алгоритм «вибору з поверненням» у вихідну вибірку, у якому формально зберігаються незмінними ступеня свободи на кожному етапі обробки даних та перетворення вибірок [6 – 10]. Він запропонував новий спосіб створення нових вибірок на основі генеральної вибірки, моделюючи вибірки з емпіричного розподілу, або, іншими словами, взяти кінцеву сукупність з n елементів вихідної вибірки і за допомогою датчика випадкових чисел сформувати з неї будь-яку кількість розмножених вибірок. Процедура,

хоч і нереальна без використання ЕОМ для перетворення вибірок, але дуже проста з погляду програмування.

Бутстреп у статистиці – практичний комп'ютерний спосіб дослідження розподілу статистик ймовірнісних розподілів, заснований на багаторазовій генерації вибірок за допомогою методу Монте-Карло з урахуванням наявної вибірки. Дозволяє просто і швидко оцінювати різні статистики для складних моделей. Суть методу бутстреп у тому, щоб у наявній вибірці побудувати емпіричний розподіл. Використовуючи цей розподіл як теоретичний розподіл ймовірностей, можна за допомогою датчика псевдовипадкових чисел згенерувати практично необмежену кількість псевдовибірок довільного розміру, наприклад, того ж, як у вихідної [10]. На безлічі псевдовибірок можна оцінити як аналізовані статистичні характеристики, а й вивчити їх імовірнісні розподіли. Таким чином, наприклад, можна оцінити дисперсію або квантили будь-якої статистики незалежно від її складності. Цей метод є методом непараметричної статистики. Поряд з методами «складного ножа», перехресної перевірки та перестановним тестуванням (англ. exact test) складає клас методів генерації повторної вибірки (англ. resampling).

Насамперед, доводиться констатувати, що підхід Б. Ефрона виник під сильним впливом ідей Р. Фішера, в основному концепції максимальної правдоподібності, що з'явилася в 1912 р. З неї, власне, випливає, що те, що спостерігалося в експерименті, якраз і повинно було статися, тому всі невідомі, які нам слід витягти з експерименту, треба знаходити таким чином, щоб вони якнайкраще узгоджувалися з наявними даними. Тоді оцінки невідомих і будуть «найбільш правдоподібними» для наявних даних. Багаторічний розвиток цієї концепції перетворив її в один із наріжних каменів сучасної математичної статистики.

Але є три обставини, які заважають нам повною мірою усвідомити переваги підходу, що базується на принципі максимальної правдоподібності [8]. Це можливе зміщення на кінцевих вибірках, потреба у суттєвій апріорній інформації (знанні виду закону розподілу досліджуваних випадкових величин)

та обчислювальні труднощі. Втім, останні не мають принципового характеру, зате з першими двома доводиться рахуватися. Бутстреп-метод спочатку виник як засіб подолання вибіркового зміщення або його істотного зменшення. Якщо з'ясується, що він коректний, то з цього випливатиме, що класична процедура методу максимуму правдоподібності не дозволяє витягти з вибірки всю наявну в ній інформацію [4].

Методи бутстреп більш гнучкі, ніж класичні методи, які можуть бути аналітично нерозв'язними або непридатними для використання через відсутність відповідних припущень.

Цей метод можна використовувати для наборів малого обсягу даних. Грубо кажучи, бутстраповська ідея наближення сукупності вибіркою викликає сумніви по мірі зменшення обсягу вибірки n [15]. Як і у випадку з іншими статистичними процедурами, наша довіра до бутстрепу зростатиме зі збільшенням розміру вибірки.

Оцінка бутстреп методом насамперед потрібна для точності експерименту, коли класичних статистичних оцінок даних не існує [6].

Метод Монте-Карло – загальна назва групи чисельних методів, заснованих на отриманні великої кількості реалізацій стохастичного (випадкового) процесу, який формується таким чином, щоб його імовірнісні характеристики збігалися з аналогічними величинами завдання, що розв'язується [1 – 4].

Історично інтенсивний розвиток теорії та додатків методу Монте-Карло був пов'язаний із розробкою чисельних моделей ядерних процесів (при створенні відповідних військових та технічних пристроїв – бомб, реакторів тощо).

Останнім часом сфера застосування чисельного статистичного моделювання значно розширилася. Було розроблено змістовну теорію імовірнісних уявлень розв'язків задач математичної фізики. За підсумками цієї теорії було побудовано ефективні (економічні) оцінювачі методу Монте-Карло.

На відміну від звичайної моделі прогнозування метод Монте-Карло передбачає набір результатів на основі передбачуваного діапазону значень, а не набору фіксованих вхідних значень. Іншими словами, моделювання методом Монте-Карло створює модель можливих результатів з використанням розподілу ймовірностей, наприклад, рівномірного або нормального розподілу для будь-якої змінної, яка містить елемент невизначеності. Потім виконується повторне обчислення результатів з іншими наборами випадкових чисел у діапазоні від мінімального до максимального значень. У типовому експерименті Монте-Карло ця операція повторюється кілька тисяч разів для створення великої кількості можливих результатів.

Крім того, висока точність методу Монте-Карло дозволяє використовувати його для довгострокового прогнозування [2]. Зі збільшенням кількості вхідних даних зростає і кількість прогнозів, що дозволяє з більшою точністю прогнозувати результати більш віддалені терміни. Результатом виконання методу Монте-Карло є діапазон можливих результатів із зазначенням ймовірності кожної події.

Як простий приклад моделювання методом Монте-Карло можна навести розрахунок ймовірності при киданні двох стандартних гральних кісток. Існує 36 можливих комбінацій двох кісток. Виходячи з цього можна вручну розрахувати ймовірність певного результату [3]. Для більш точного прогнозу можна 10000 разів повторити кидок кісток, використовуючи метод Монте-Карло.

Незалежно від інструменту, метод Монте-Карло складається з трьох основних кроків.

1. Створення прогнозної моделі з визначенням залежної змінної, щодо якої здійснюється прогноз, та незалежних змінних (також відомих як вхідні дані, змінні ризики або передикторні змінні), що лежать в основі прогнозу.

2. Визначення розподілу ймовірностей незалежних змінних [3]. Визначення діапазону ймовірних значень за допомогою наявних статистичних

даних та/або суб'єктивних знань аналітика з наступним присвоєнням кожному такому значенню вагових коефіцієнтів ймовірності.

3. Багаторазове виконання моделювання створення випадкових значень незалежних змінних. Моделювання виконується до того часу, поки буде отримано репрезентативна вибірка практично нескінченного числа можливих комбінацій.

Змінюючи базові параметри моделювання, метод Монте-Карло можна повторювати скільки завгодно разів. Також для обчислення розкиду у вибірці можна розрахувати дисперсію та стандартне відхилення, які традиційно використовуються для оцінки розкиду [5]. Як правило, чим менша дисперсія, тим краще.

Метод Монте-Карло може бути застосований для оцінки невизначеності фінансових прогнозів, результатів інвестиційних проектів, при прогнозуванні вартості та графіка виконання проекту, порушень бізнес-процесу та заміни персоналу. Даний метод застосовують у ситуаціях, коли результати не можуть бути отримані аналітичними методами або існує висока невизначеність вхідних чи вихідних даних.

1.2 Чотири алгоритми бутстрепу

Як вже згадувалося раніше – основна ідея бутстрепу полягає у тому, щоб методом статистичних випробувань Монте-Карло багаторазово отримувати повторні вибірки з емпіричного розподілу. Бутстреп, як і інші методи генерації повторних вибірок, корисні, коли значення статистичних характеристик не можна отримати з теоретичних припущень через недостатній обсяг даних вибірок.

Залежно від наявної інформації розрізняють параметричний та непараметричний бутсреп.

Непараметрична бутсреп-процедура складається з отримання великої кількості повторностей (випадкових наборів із сукупності) з одної випадкової вибірка, отриманої емпіричним шляхом.

Замість того, щоб робити нові повторності експерименту, на основі однієї вибірки генерується безліч псевдовибірок того ж розміру, що складаються з випадкових комбінацій вихідного набору елементів [16]. У цьому використовується алгоритм «випадкового вибору із поверненням» (random sampling with replacement).

Якби здійснювався метод без повернення (random sampling without replacement), то завжди отримували б вихідну безліч чисел, представлену щоразу у різному порядку.

Наступний крок непараметричної процедури: побудова бутстреп-розподілу величини, що оцінюється. Для кожної псевдоповторності, отриманої на кроці 1, розраховується значення аналізованої характеристики – вибіркового середнього, медіани, стандартного відхилення та ін.

Параметричний бутстреп використовує припущення, що вихідні вибіркові дані є випадковими реалізаціями ймовірнісного процесу, що визначається деяким теоретичним розподілом.

Процедура параметричного бутстрепу складається з кількох кроків. У першому з них за вибірковими даними $\{x_1, x_2, ..., x_n\}$ здійснюється побудова імовірнісної моделі та оцінюються її параметри θ (наприклад, математичне сподівання і середнє квадратичне відхилення у випадку нормального закону розподілу). Наступний крок полягає у тому, що з випадковим чином підібраного розподілу з параметрам $\theta\hat{}$ генерується n елементів $\{x_1^*, x_2^*, ..., x_n^*\}$, і бутстреп-повторність, отримана такою імітацією, використовується для розрахунку значення статистики $t^* = T(x^*)$. При виконанні третього кроку другий крок повторюється B раз і формується бутстреп-розподіл аналізованої статистики $\{t_1, t_2, ..., t_j, ..., t_B\}$ [16].

У загальному вигляді непараметрична бутстреп-процедура виглядає так.

Крок 1. Отримання великої кількості повторностей – випадкових наборів даних із сукупності. В якості вихідних даних береться, як правило, тільки одна випадкова вибірка, отримана емпіричним шляхом. Замість того щоб робити нові повторності експерименту, на основі однієї наявної вибірки генерується безліч псевдовибірок того ж розміру, що складаються з випадкових комбінацій вихідного набору елементів. У цьому використовується алгоритм «випадкового вибору із поверненням», тобто витягнуте число знову поміщається у «колоду, що перемішується» перш ніж витягується наступне спостереження. У результаті деякі члени у кожній окремій псевдовибірці можуть повторюватися два або більше разів, тоді як інші відсутні. Зазначимо, що якби ми здійснювали вибір без повернення, то весь час отримували б вихідну множину чисел, хоч і представлену щоразу в різному порядку.

Крок 2. Побудова бутстреп-розподілу величини, що оцінюється. Для кожної псевдоповторності, отриманої на кроці 1, розраховується значення аналізованої характеристики – вибіркового середнього, медіани, стандартного відхилення та ін. Маючи таку множину даних, легко побудувати гістограму (або згладжений графік щільності частотного розподілу) значень показника, що відображає закономірності його варіації і дає можливість оцінити довірчі інтервали та інші корисні вибіркові характеристики аналізованої величини. Якщо простий непараметричний бутстреп виконує перевибірку з урахуванням рівної ймовірності появи кожного елемента, стратифікований бутстреп враховує співвідношення частот між відносно гомогенними групами (стратами), на які може бути розділені вибіркові об'єкти.

Мета аналізу даних – отримати максимально точні вибіркові оцінки та поширити результати на всю популяцію.

Одночасно з впровадженням методів планування експерименту почали бурхливо розвиватися алгоритми рандомізації, які полягають у багаторазовому випадковому перемішуванні рядків або стовпців таблиці

спостережень щодо рівнів впливу факторів, що вивчаються. При кожній ітерації перестановочного тесту на основі згенерованої псевдовибірки розраховуються імітовані значення аналізованого показника або статистики, які порівнюються з аналогічною величиною, знайденою за емпіричними даними. У результаті перестановок не змінюється ні склад вихідної таблиці, ні чисельність груп із різними рівнями впливу, лише відбувається безладний обмін елементами даних між цими групами.

Існують думки, що рандомізація взагалі є окремим випадком випробувань Монте-Карло. Однак, незважаючи на схожість цих методів в основних алгоритмах і обмеженнях, між ними є досить суттєві концептуальні відмінності (наприклад, для методів Монте-Карло типові дослідження, коли дані спостережень взагалі не використовуються, щоб змоделювати ймовірнісний процес) [14].

Процедури ресамплінгу не вимагають жодної апріорної інформації про закон розподілу випадкової величини, що вивчається, і в цьому сенсі можуть розглядатися як непараметричні. Вони виконують обробку різних фрагментів вихідного масиву емпіричних даних, хіба що повертаючи їх «різними гранями» і зіставляючи отримані таким чином результати. Питання про повну коректність такого прийому залишається відкритим, але якщо визнати його законним, то асимптотичні переваги ресамплінгу в порівнянні з класичними параметричними тестами стають очевидними. Значення параметрів, побудованих за розмноженими підвиборками, строго кажучи, не є незалежними, проте при збільшенні n з ресемпльованими значеннями статистик, можна поводитися як з незалежними випадковими величинами [13].

Ключовим є обсяг вибірки. Що робити, якщо обсяг вибірки невеликий? Один з розумних підходів полягає в тому, щоб випадково витягувати дані з наявної вибірки. Зазвичай випадковим чином генерується кілька тисяч вибірок, з цього набору можна знайти бутстреп-розподіл статистики, що нас цікавить.

Для кожної вибірки будується оцінка шуканої величини, далі оцінки усереднюються. Оскільки вибірок багато, можна побудувати емпіричну функцію розподілу оцінок, далі розрахувати квантиль, обчислити довірчий інтервал.

Ясно, що бутстреп метод є модифікацією методу Монте-Карло.

Ідея бутстрепу полягає у тому, щоб використовувати результати обчислень за вибірками як «фіктивну популяцію» з метою визначити вибірковий розподіл статистики. Фактично, аналізується велика кількість «фантомних» вибірок, званих бутстреп-вибірками.

Як працює оцінка методом бутстрепу? Алгоритм наступний. Нехай ми маємо ряд точок даних від от $x_1, x_2, \ldots, x_n$. Ми беремо вибірку з поверненням x_b з цих даних B разів, кожен розмірністю N. Для кожної з B підвибірок обчислюється шуканий параметр – це може бути середнє вибіркове, дисперсія або будь-який інший статистичний показник.

Після того, як цикл буде закінчено, у нас виявиться B різних оцінок даного параметра, які можна використовувати для знаходження середнього значення та дисперсії параметра. Навіщо нам потрібні середнє вибіркове значення та дисперсія? Насамперед середнє значення показує найбільш ймовірне значення параметра чи, іншими словами, очікуване значення параметра, а дисперсія показує точність такої оцінки. Велика дисперсія означає, що середнє значення є неточним, мала дисперсія означає більшу точність.

Вибірка з поверненням.

Припустимо, ми маємо набір даних {1, 2, 3, 4, 5}. Припустимо також, що ми вибираємо один із цих елементів і це виявляється 5. Вибірка з поверненням означає, що, якщо ми виберемо елемент ще раз, ми можемо знову отримати 5. Насправді можна набрати підвибірку з одних п'ятірок, оскільки ми повертаємо елемент після того як взяли його з набору даних. Протилежним випадком є вибірка без повернення. У разі, якщо ми виберемо число елементів, що

дорівнює набору даних, ми просто виберемо сам набір даних. Тому нам важлива саме вибірка із поверненням.

Метод бутстрепа полягає в тому, що одну реальну вибірку з генеральної сукупності за допомогою електронних обчислювальних машин тиражують у великій кількості екземплярів, а потім з отриманого масиву випадковим чином роблять необхідне (дуже велике) число нових вибірок, які потім аналізують.

Коли частки вибірки на першому етапі малі, методи бутстрепа для стратифікованих багатоступінчастих планів вибірки спрощуються, оскільки вибірка без заміни може бути апроксимована вибіркою із заміною. Але коли фракціями вибірки на першому етапі не можна знехтувати, методи бутстреп для узгодженої оцінки дисперсії стають складними, і їх мало розроблено [16]. Наприклад, для стратифікованої триетапної вибірки з простою випадковою вибіркою без заміни кожної стадії довільними фракціями вибірки не доступна процедура бутстреп, крім бутстрепу Бернулли (Bernoulli Bootstrap), запропонованої Фунаока, Сайго, Ситтер і Тойда для Національного огляду цін (NSP) Японії 1997 року.

1.3 Бутстреп Бернуллі

Бутстреп-вибірка будується шляхом випадкової заміни вибраних одиниць. Процедура проводиться самостійно для $h = 1, 2, \ldots, H$.

Крок 1. Вибрати ($n_h = 1$) міток простою випадковою вибіркою із заміною з S_{h1}. Позначимо кандидата, встановленого C^*_{h1}. Для кожного $i \in S_{h1}$ ми: (a) зберігаємо його в бутстраповій вибірці з ймовірністю $p_{h1} = 1 - (1/2)(1 - n_h^{-1})^{-1}(1 - f_{1h})$; або (b) замінити його випадковим чином вибраним із C^*_{h1}.

Крок 2. Для і, збереженого на кроці 1, виберіть $(m_{hi} - 1)$ мітки простою випадковою вибіркою із заміною S_{h2i}. Позначимо кандидата, встановленого C^*_{h2i}. Для кожного j S_{h2i} ми: (c) зберігаємо його у вибірці бутстрапу з

ймовірністю $p_{h2i} = 1 – (½)p_h^{-1}f_{1h}(1 – m_{mi}^{-1})^{-1}(1 – f_{2hi})$; або (d) замінити його випадковим чином обраним C^*_{h2i}. Якщо це так, перейдіть до кроку 3.

Крок 3. Для j, збереженого на кроці 2, виберіть $(l_{hij} – 1)$ мітки простою випадковою вибіркою із заміною S_{h3ij}. Позначимо кандидата, встановленого C^*_{h3ij}. Для кожного $k \in S_{h3ij}$ ми: (e) зберігаємо його в бутстраповій вибірці з ймовірністю $p_{h3ij} = 1 – (1/2)p_h^{-1}f_{1h}p_{h2i}^{-1}f_{2hi}(1 – l_{hij}^{-1})^{-1}(1 – f_{3hij})$; або (f) замінити його одним, випадково вибраним із C^*_{h3ij}.

Позначимо отриманий бутсреп як $\{S^*_{h1}, h = 1, 2, \ldots, H; S^*_{h2i}, i \in S_{h1}; S^*_{h3ij} \in S^*_{h2i}\}$ і нехай $w^*_{hijk} = w_{hijk}$.

Створення наборів кандидатів (candidate) необхідно, щоб зробити процедуру здійсненною для будь-яких $n_h, m_{hi}, l_{hij} \geq 2$.

Зберігаються вихідні розміри вибірки та вихідні ваги вибірки. Це бажано під час роботи з випадково розрахованими даними обстежень.

Байєсовський бутстреп.

Припустимо, у нас є вибірка розміром n, скажімо $x_1, \ldots, x_n$, яка розглядається як n незалежних та однаково розподілених реалізацій випадкових величин X. Статичний $\varphi`$ обирається для оцінки параметра $\varphi$ розподілу $X$; $x_1$, $X$, $\varphi`$ і φ можуть бути векторами. Розподіл бутстреп для φ генерується повторним копіюванням бутсреп $x_1, x_2, \ldots, x_n$.

Одне повторне копіювання бутстреп з $x_1, x_2, \ldots, x_n$ є простою випадковою вибіркою розміру n із $x_1, x_2, \ldots, x_n$ з поверненням, а одне повторне копіювання бутстреп $\varphi`$ – це значення $\varphi$, обчислене на повторній копії бутстрепу. Бутстреп розподіл $\varphi`$ генерується з урахуванням усіх можливих повторних копій $\varphi`$.

Бутстреп регресії гаусівського процесу.

Коли дані корелюються у часі, пряма бутстреп руйнує внутрішні кореляції. У цьому методі використовується регресія гаусівського процесу, щоб відповідати ймовірності моделі, з якої потім можуть бути побудовані репліки. Регресія гаусівського процесу – це метод байєсівської нелінійної

регресії. Гаусівський процес – це набір випадкових величин, будь-яке кінцеве число яких має спільний гауссівський (нормальний) розподіл. Гаусівський процес визначається функцією середнього та функцією коваріації, які визначають вектори середніх значень та матриці коваріації для кожного кінцевого набору випадкових величин.

1.4 Побудова довірчих інтервалів

З'ясуємо, які статистики краще використовувати для побудови довірчих інтервалів за допомогою бутстрепа. По-перше, бутстрепівський розподіл є центрованим не навколо справжнього значення статистики, а навколо його вибіркового аналога. По-друге, слід бутстрепувати асимптотично пивотальні статистики.

Розглянемо кілька варіантів бутстрепівських статистик, що використовуються для побудови довірчих інтервалів та підкреслимо їх позитивні та негативні якості. Нехай нас цікавить побудова статистичних висновків відносно параметра β із її оцінки $\hat{\beta}$.

Ефронівський довірчий інтервал. У даному випадку статистикою, яка оцінюється бутстреп-процедурою, є сама оцінка, тобто, $\hat{\theta} = \hat{\beta}$. Таким чином, ми отримуємо бутстрепівський розподіл $\left\{\hat{\theta}_b^* = \hat{\beta}_b^*\right\}_{b-1}^{B}$. Відповідні квантилі розподілу $q_{\alpha/2}^*, q_{1-\alpha}^*$, довірчий інтервал $CI_E = \left[q_{\alpha/2}^*, q_{1-\alpha}^*\right]$.

Ефронівський довірчий інтервал був популярним, коли бутстрепівський підхід тільки починав використовуватися. Насправді цей довірчий інтервал дає непогану апроксимацію для справжніх рівнів значущості, оскільки зберігає зміщення вихідної вибірки.

Холівський довірчий інтервал. Холл запропонував використовувати для побудови довірчого інтервалу рецентровану статистику $\hat{\theta} = \hat{\beta} - \beta$, що знімає проблему усунення, пов'язаного з кінцівкою вибірки. Таким чином,

отримується бутстрепівський розподіл $\left\{\hat{\theta}_b^* = \hat{\beta}_b^* - \hat{\beta}\right\}_{b-1}^{B}$. Відповідні квантилі розподілу $q_{\alpha/2}^*, q_{1-\alpha}^*$, довірчий інтервал $CI_H = \left[\hat{\beta} - q_{1-\alpha/2}^*, \hat{\beta} - q_{\alpha/2}^*\right]$.

Холівський довірчий інтервал дає кращу апроксимацію рівнів значущості, ніж Ефронівський. Плюсом використання Холлівського довірчого інтервалу є відсутність необхідності оцінювання стандартних помилок.

Також існує *t* - відсотковий довірчий інтервал. Такий інтервал використовують як статистику, що бутстрепується, *t*-статистику, тобто $\left(\hat{\beta} - \beta\right) / se\left(\hat{\beta}\right)$. Таким чином, знаходять бутстрепівський розподіл статистики $\left\{\frac{\hat{\beta}_b^* - \hat{\beta}}{se\left(\hat{\beta}_b^*\right)}\right\}_{b=1}^{B}$ і відповідні квантилі $q_{\alpha/2}^*, q_{1-\alpha}^*$, а саме *t* - відсотковий довірчий інтервал будують як

$$CI_t = [\hat{\beta} - se(\hat{\beta}) q_{1-\alpha/2}^*, \hat{\beta} - se(\hat{\beta}) q_{\alpha/2}^*].$$

Зауважимо, що *t*-відсотковий довірчий інтервал набагато краще апроксимує справжні рівні значущості, ніж Холівський довірчий інтервал. Але використовувати його рекомендується у випадках, коли стандартні помилки можна побудувати якісно.

Симетричний *t*-відсотковий довірчий інтервал. Такий інтервал використовує як бутстрепований "симетризовану" t-статистику $\frac{|\hat{\beta} - \beta|}{se(\hat{\beta})}$.

Розподіл бутстрепівської статистики є $\left\{\frac{|\hat{\beta}_b^* - \hat{\beta}|}{se(\hat{\beta}_b^*)}\right\}_{b=1}^{B}$, а правий кванитиль – q_α^*.

Симетричний *t* -відсотковий довірчий інтервал:

$$CI_{|t|} = [\hat{\beta} - se(\hat{\beta}) q_{1-\alpha/2}^*, \hat{\beta} + se(\hat{\beta}) q_{1-\alpha/2}^*].$$

Симетричний t-відсотковий довірчий інтервал має у певних випадках перевагу перед t-відсотковим довірчим інтервалом. А саме, якщо асимптотичний розподіл статистики $\hat{\beta} - \beta$ є симетричним (у разі асимптотичної нормальності), то CI_t дає кращу апроксимацію рівнів значущості.

Довірчі інтервали, засновані на модифікованому бутстрепі.

Основна ідея методу бутстрепу базується на тому факті, що емпіричний розподіл F_n зближується із дійсним розподілом F_o. Звідси можна зробити висновок, що розподіл оцінок, отриманих за спостереженнями, що мають розподіл F_n, близький до розподілу оцінок за спостереженнями з розподілом F_o.

З іншого боку, довірчі інтервали мають таку властивість: кінці інтервалу відповідають розподілам, що досить сильно розрізняються між собою. Іншими словами, можливість отримати вибірку з «сильно розрізненими» спостереженнями дорівнює α, тобто, одиниця мінус рівень довіри.

Наша основна ідея полягає у побудові двох розподілів $Fn+$ і $Fn-$ таких, що ймовірність отримати оцінку, що спостерігається θ_n, дорівнює α. Ми також вимагаємо, щоб ці два розподіли були побудовані за вибіркою, що спостерігається. Крім цих обмежень, ми ще вимагаємо, щоб ці розподіли були близькі до емпіричного розподілу.

Іншими словами, ми хочемо змінити функцію емпіричного розподілу, але не більше, ніж це необхідно. Довірчі інтервали будуть визначатися значеннями θ, відповідними $Fn-$ і $Fn+$.

Поява залежних спостережень впливає на ймовірнісні властивості емпіричного розподілу. Ми будемо вимірювати цей вплив, використовуючи такий лінійний вираз, який поводить себе як псевдо-значення:

$$z_i = \sum_{j=1}^{n} \theta(x_1,...,x_{j-1},x_i,x_{j+1},...,x_n) - (n-1)\theta_n$$

Якщо θ визначено для вибірки обсягом (*n-1*), останній вираз може бути замінено звичайним псевдо-значенням з теорії статистик з використанням методу складаного ножа. Розподіли *Fn* – і *Fn*+ обираються у вигляді $p(x_i) \infty \exp\left(c\left(z_i - \bar{z}\right)\right)$, де *c* обрано так, щоб ймовірності «хвостів» були б точно рівні α і (1- α) (знак ∞ означає пропорційність лівої і правої частин, $\bar{z} = \sum_{i=1}^{n} z_i$).

Якщо θ лінійна по z_i, ці розподіли мають найбільшу ентропію серед усіх розподілів з такими ж «хвостовими» ймовірностями. Апроксимація довірчого інтервалу для *θ(F)* може бути побудована у вигляді інтервалу $(\theta^-, \theta^+) = (\theta(F^-), \theta(F^+))$.

Легко здійснити повторну вибірку з розподілів *Fn*+ і *Fn* – тим самим способом, що й повторний вибір у звичайному бутстрепі. Візьмемо *N* незалежних бутстреп вибірок із простим випадковим вибором із поверненням із спостережуваної вибірки. Для кожної бутстреп вибірки обчислимо відповідні оцінки $\theta^*_{n,i}$ *i = 1, 2, …, n.*

Для фіксованого значення з «хвостовими» ймовірностями відповідні розподіли можуть бути легко оцінені виразом:

$$P^*(\theta^* \le a \mid c) = \frac{\sum_{j:\theta^*_n, j \le a} \exp(c \sum (z_j, i - \bar{z}))}{(\sum_j \exp(c \sum (z_j, i - \bar{z})))}$$

Використовуючи інтерполяцію, легко отримати значення, що дають коректні «хвостові» ймовірності для $a = \theta_n$ (тобто α і (1- α)). Оскільки визначено два значення *c*, то відповідні значення дають бутстреп інтервал θ^- і θ^+.

1.5 Тестування гіпотез за допомогою бутстрепу

Однією з основних цілей бутстрепу є тестування гіпотез. Розглянемо, як за допомогою бутстрепу тестуються найпростіші статистичні гіпотези. Нехай нульова гіпотеза має вигляд $H_0 : \beta = \beta^0$, де $\beta -$ скаляр.

Альтернативна гіпотеза є односторонньою: $H_\alpha : \beta \succ \beta^0$. Бутстрепуємо t -відсоткову статистику:

$$\hat{\theta} = \frac{\hat{\beta} - \beta}{se(\hat{\beta})}.$$

Як результат отримуємо бутстрепівський розподіл цієї статистики і відповідний квантиль:

$$\left\{ \hat{\theta}_b^* = \frac{\hat{\beta}_b^* - \hat{\beta}}{se(\hat{\beta}_b^*)} \right\}_{b=`}^{B} \Rightarrow q_{1-\alpha}^* .$$

Гіпотеза H_0 відкидається, якщо $\frac{\hat{\beta} - \beta^0}{se(\hat{\beta})} > q_{1-\alpha}^*$.

Альтернативна гіпотеза є двосторонньою : $H_\alpha : \beta \neq \beta^0$. У такому випадку ми будемо бутстрепити симетричну t -процентну статистику:

$$\hat{\theta} = \frac{|\hat{\beta} - \beta|}{se(\hat{\beta})}.$$

Отримуємо бутстрепівський розподіл та квантиль:

$$\left\{ \hat{\theta}_b^* = \frac{|\hat{\beta}_b^* - \hat{\beta}|}{se(\hat{\beta}_b^*)} \right\}_{b=`}^{B} \Rightarrow q_{1-\alpha}^* .$$

Гіпотеза H_0 відкидається, якщо $\frac{|\hat{\beta}-\beta^0|}{se(\hat{\beta})} > q^*_{1-\alpha}$.

Нехай нульова гіпотеза має вигляд $H_0 : \beta = \beta^0$, де β – вектор. У такому випадку ми будуємо Вальдовську статистику (з точністю до коефіцієнта пропорційності):

$$\hat{\theta} = (\hat{\beta}-\beta)' \hat{V}_\beta^{-1} (\hat{\beta}-\beta).$$

Відповідно, отримуємо бутстрепівський розподіл та квантиль:

$$\left\{ \hat{\theta}_b^* = (\hat{\beta}_b^* - \hat{\beta})' \hat{V}_\beta^{*-1} (\hat{\beta}_b^* - \hat{\beta}) \right\}_{b=1}^{B} \Rightarrow q^*_{1-\alpha}$$

Гіпотеза H_0 відкидається, якщо $\hat{\theta}^0 = (\hat{\beta}-\beta^0)' \hat{V}_\beta^{-1} (\hat{\beta}-\beta^0) > q^*_{1-\alpha}$.

Нехай тепер нульова гіпотеза має вигляд лінійних обмежень на коефіцієнти $H_0 : R\beta = r$, де R – матриця обмежень. У такому випадку ми знову бутстрепуємо Вальдовську статистику (з точністю до коефіцієнта пропорційності):

$$\hat{\theta} = (R\hat{\beta}-r)' (R\hat{V}_\beta R')^{-1} (R\hat{\beta}-r).$$

Отримуємо бутстрепівський розподіл, з якого знаходимо відповідний квантиль:

$$\left\{ \hat{\theta}_b^* = (\hat{\beta}_b^* - \hat{\beta})' R' (R\hat{V}_\beta^* R')^{-1} R(\hat{\beta}_b^* - \hat{\beta}) \right\}_{b=1}^{B} \Rightarrow q^*_{1-\alpha}.$$

Зауважимо, що ми рецентруємо бутстрепівську статистику. Без цього

бутстрепівський розподіл успадкував би усунення, властиве початковій статистиці. Гіпотеза H_0 відкидається, якщо

$$\hat{\theta} = (R\hat{\beta} - r)'(R\hat{V}_{\beta}R')^{-1}(R\hat{\beta} - r) > q^*_{1-\alpha}.$$

Асимптотичне рафінування.

Іноді кажуть, що за допомогою бутстрепу досягається асимптотичне рафінування. Розглянемо, що таке асимптотичне рафінування і у яких випадках воно має місце.

Нехай є певна статистика $\hat{\theta}$, істинний розподіл якої $F_{\hat{\theta}}(x)$. Позначимо бутстрепівський розподіл цієї статистики через $F^*_{\hat{\theta}}(x)$.

Кажуть, що за допомогою бутстрепу досягається асимптотичне рафінування, якщо помилка апроксимації істинного розподілу $F_{\hat{\theta}}(x)$ бутстрепівським $F^*_{\hat{\theta}}(x)$ більшого порядку малості, ніж помилка апроксимації асимптотичним розподілом при прагненні обсягу вибірки до нескінченності.

Наведемо приклади, що використовують розкладання Еджворта функції розподілу статистики навколо граничного розподілу.

Асимптотично пивоталья t-статистика. Нехай статистика, що бутстрепується $\hat{\theta} = \dfrac{\hat{\beta} - \beta}{se(\hat{\beta})}$. Її асимптотичний розподіл є стандартним нормальним: $\hat{\theta} \xrightarrow{d} N(0,1)$ (тобто статистика асимптотично пивотальна). Позначимо точний розподіл статистики через $F_{\hat{\theta}}(x)$, а бутстрепівський – через $F^*_{\hat{\theta}}(x)$. Для кумулятивної функції стандартного нормального розподілу використовуємо звичайне позначення $\Phi(x)$.

Отже, розкладемо справжній та бутстрапівський розподіли навколо асимптотичного:

$$F_{\hat{\theta}}(x) = \Phi(x) + \frac{h_1(x,F)}{\sqrt{n}} + \frac{h_2(x,F)}{n} + O\left(\frac{1}{n^{3/2}}\right).$$

Тут $h_1(x,F)$ – парна по x безперервна F функція, $h_2(x,F)$ – непарна по x, безперервна F функція. Помилки апроксимації точного розподілу асимптотичним та бутстреповським, відповідно, дорівнюють:

$$\Phi(x) - F_{\hat{\theta}}(x) = \frac{h_1(x,F)}{\sqrt{n}} + O\left(\frac{1}{\sqrt{n}}\right) = O\left(\frac{1}{\sqrt{n}}\right);$$

$$F^*_{\hat{\theta}}(x) - F_{\hat{\theta}}(x) = \frac{h_1(x,\hat{F}) - h_1(x,F)}{\sqrt{n}} + O\left(\frac{1}{n}\right) = O\left(\frac{1}{n}\right).$$

Тут скористалися тим фактом, що різниця $h_1(x,\hat{F}) - h_1(x,F)$ має асимптотику $\frac{1}{\sqrt{n}}$ так як

$$\sqrt{n}\left(\hat{F}(x) - F(x)\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} 1[x_i \le x] - E[1[x_i \le x]]\right)$$

$$\xrightarrow{d} N(0, P\{x_i \le x\}P\{x_i > x\}).$$

Таким чином, у даному прикладі використання бутстрепу призводить до асимптотичного рафінування.

Асимптотично непівотальна статистика. Розглянемо статистику

$$\theta = \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$$

При збереженні позначення кумулятивних функцій розподілу для точного розподілу та бутстрепівського з попереднього пункту, позначимо асимптотичний розподіл через $\Phi(x, V_\beta)$.

Зауважимо, що у такому випадку статистика асимптотично непивотальна, тобто. її асимтотичний розподіл залежить від невідомого параметра, у даному разі V_β. Як у попередньому прикладі, розкладемо точний та бутстрепівський розподіли навколо асимптотичного:

$$F_{\hat{\theta}}(x) = \Phi(x, V_\beta) + \frac{h_1(x,F)}{\sqrt{n}} + O\left(\frac{1}{n}\right);$$

$$F^*_{\hat{\theta}}(x) = \Phi(x, V^*_\beta) + \frac{h_1(x,\hat{F})}{\sqrt{n}} + O\left(\frac{1}{n}\right).$$

Помилки апроксимації для асимптотичного та бутстрепівського розподілів вважаються аналогічними попередньому прикладу:

$$\Phi(x, V_\beta) - F_{\hat{\theta}}(x) = -\frac{h_1(z,F)}{\sqrt{n}} + O\left(\frac{1}{n}\right) = O\left(\frac{1}{\sqrt{n}}\right),$$

$$F^*_{\hat{\theta}}(x) - F_{\hat{\theta}}(x) = \Phi(x, V^*_{\hat{\theta}}) - \Phi(x, V_{\hat{\theta}}) + O\left(\frac{1}{n}\right) = O\left(\frac{1}{\sqrt{n}}\right).$$

Як видно, у даному випадку використання бутстрепу не призводить до асимптотичного рафінування. Взагалі, як правило, бутстрепування асимптотично непивотальних статистик не дає асимптотичного рафінування.

Асимптотично пивотальна симетрична t-статистка. Тепер розглянемо як приклад симетричну t –статистику:

$$\hat{\theta} = \frac{|\hat{\beta} - \beta|}{\hat{se}(\beta)} \xrightarrow{d} |N(0,1)|.$$

Зберігаючи позначення попередніх прикладів, розкладемо точний та бутстрепівські розподіли:

$$F_{\hat{\theta}}(x) = \Pr\{-x \le \frac{\hat{\beta} - \beta}{\hat{se}(\beta)} \le x\} = 2\Phi(x) - 1 + \frac{2h_2(x,F)}{n} + O\left(\frac{1}{n^{3/2}}\right),$$

$$F^*_{\hat{\theta}}(x) = 2\Phi(x) - 1 + \frac{2h_2(x,\hat{F})}{n} + O\left(\frac{1}{n^{3/2}}\right).$$

Таким чином, помилки апроксимації для асимптотики та бутстрепу мають порядки:

$$2\Phi(x) - 1 - F_{\hat{\theta}}(x) = O\left(\frac{1}{n}\right);$$

$$F^*_{\hat{\theta}}(x) - F_{\hat{\theta}}(x) = \frac{2}{n}\left(h_2(x,\hat{F}) - h_2(x,F)\right) + O\left(\frac{1}{n^{3/2}}\right) = O\left(\frac{1}{n^{3/2}}\right).$$

Таким чином, отримуємо асимптотичне рафінування. Зауважимо, що бутстрепування симетричного двостороннього тесту має помилку вищого порядку, ніж бутстрепування одностороннього тесту.

2 АНАЛІЗ ДАНИХ ЗАБРУДНЮВАЛЬНИХ РЕЧОВИН З ВИКОРИСТАННЯМ БУТСТРЕП-ПРОЦЕДУРИ

Для ілюстрації бутстреп методу були використані дані спостережень на мережі стаціонарних постів міста Запоріжжя за 2011 – 2013 рр. за основними забруднювальними речовинами (формальдегід, пил, фенол).

Дослідження виконано шляхом застосування методу Bag of Little Bootstraps (BLB). Усі дані було згруповано у єдину вибірку, з якої потім будувалися випадковим чином підвибірки різного обсягу. Розрахунки проводилися за наступною схемою:

1) з експериментальних даних випадковим чином утворювали 1000 підвибірок однакового обсягу;

2) для кожної підвибірки застосовували бутстреп метод з формуванням 1000 вибірок бутстрепа та обчисленням вибіркового середнього;

3) 1000 бутстреп характеристик вибіркового середнього ранжували за збільшенням та будували 95% довірчий інтервал;

4) кроки 1)-3) повторювали для вибірок обсягом від 10 до 100 з шагом 10.

На рисунках 2.1 – 2.3 представлені гістограми частот для значень концентрації пилу, фенолу та формальдегіду. Перевірка даних на розподіл за нормальним законом виконана за допомогою критерію Колмогорова-Смірнова. Дані не підпорядковуються нормальному розподілу (табл.2.1).

Результати обчислень вибіркового середнього представлені на рисунках 2.4, 2.5 у вигляді діаграми розмаху для вибірок обсягом від 10 до 100 елементів та рисунку 2.6 – для вибірок від 10 до 70 елементів. На діаграмах прямокутник містить значення між першим і третім квартилями, вертикальні вуса відображають значення за межами верхнього й нижнього квартилів, рисками позначені мінімальне та максимальне значення, кругами – викиди. Лінія у

прямокутнику є медіаною, тобто таким значенням ознаки, яке розташоване всередині ряду розподілу.

Таблиця 2.1 – Характеристики забруднювальних речовин

	середнє вибіркове	Значення критерію Колмогорова-Смірнова
Пил	0.4492	p-value < 2.2e-16
Фенол	0.01095	p-value < 2.2e-16
Формальдегід	0.0136	p-value = 5.793e-15

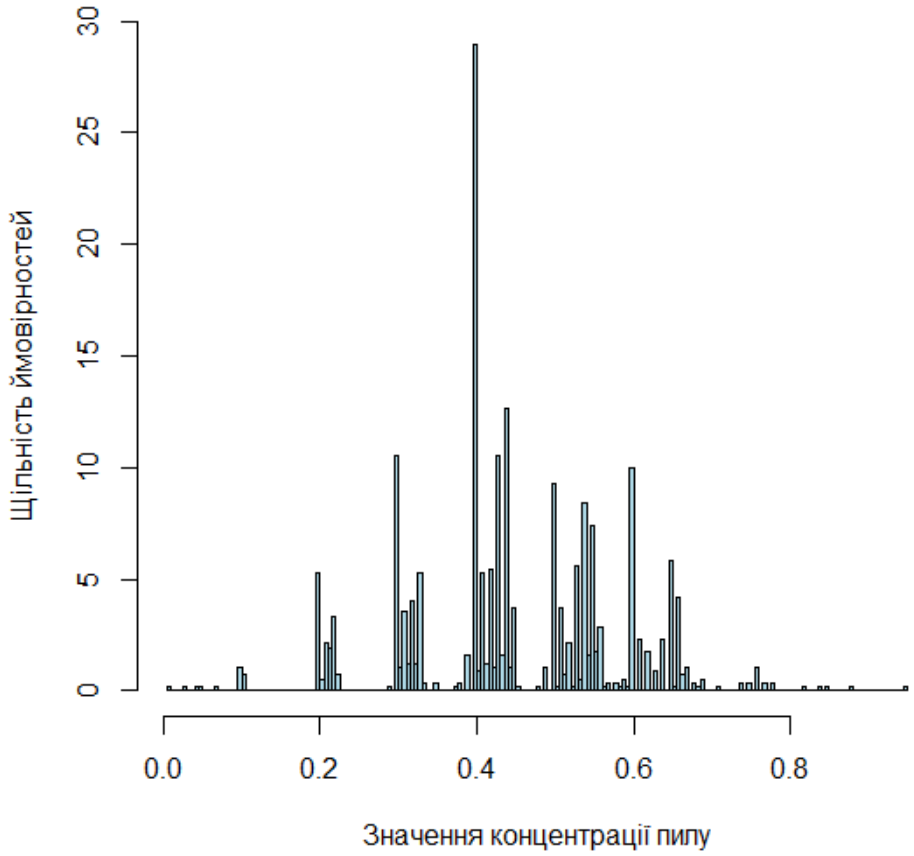

Рисунок 2.1 – Гістограма частот для значень концентрації пилу

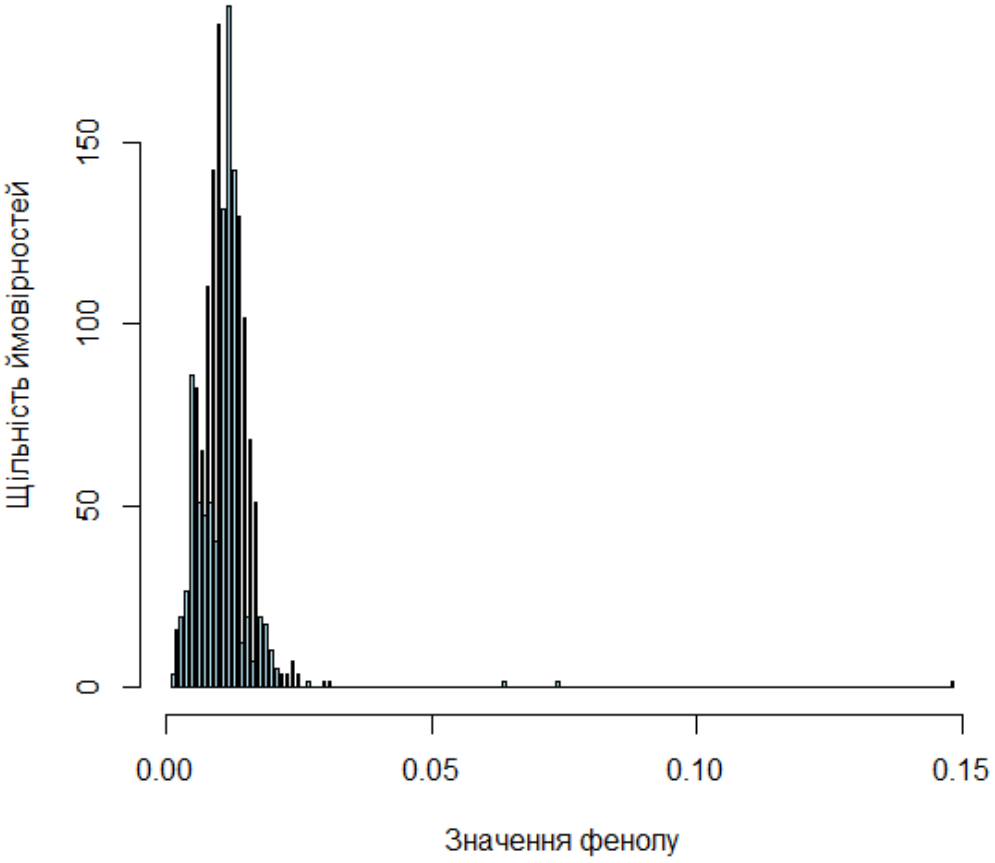

Рисунок 2.2 – Гістограма частот для значень фенолу

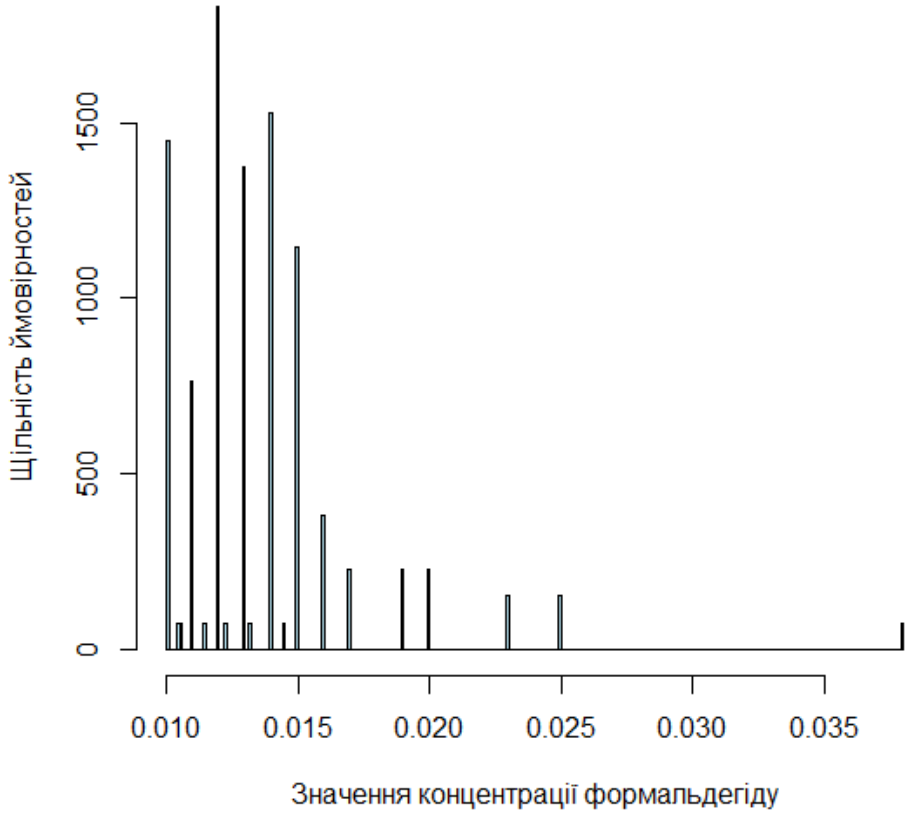

Рисунок 2.3 – Гістограма частот для значень формальдегіду

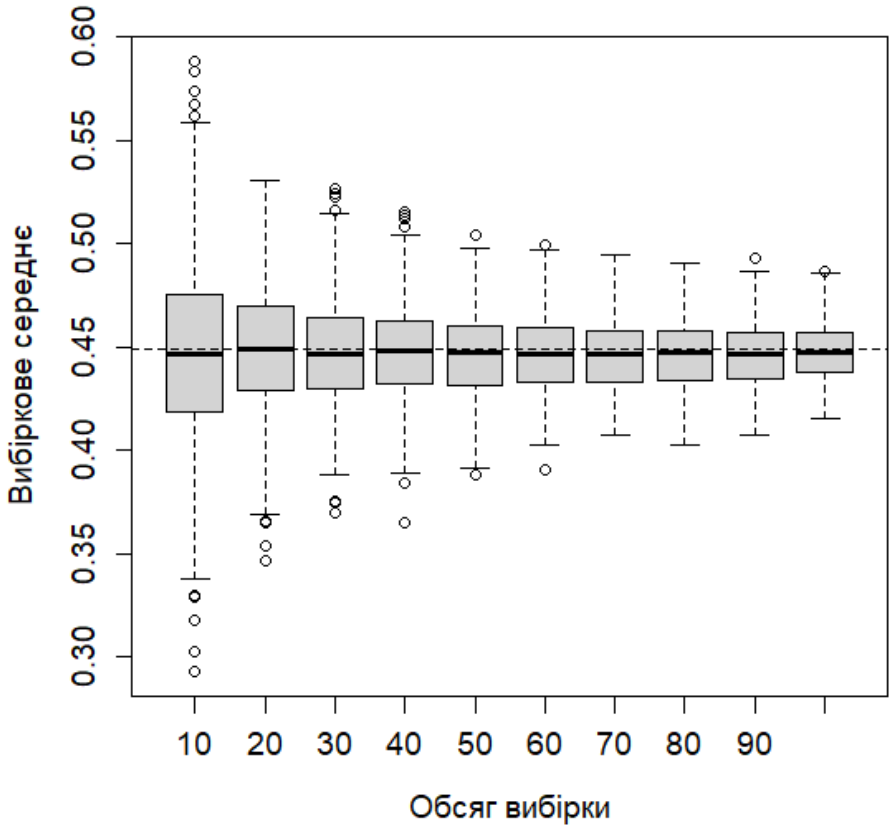

Рисунок 2.4 – Розподіл вибіркового середнього значень концентрації пилу

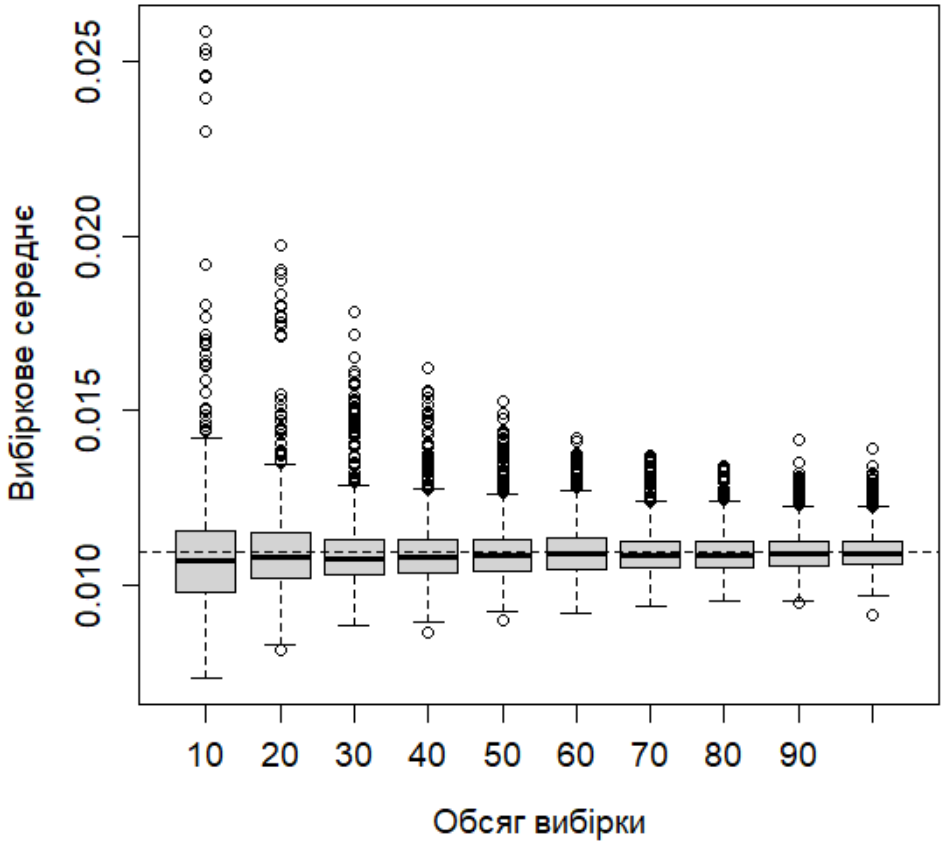

Рисунок 2.5 – Розподіл вибіркового середнього значень фенолу

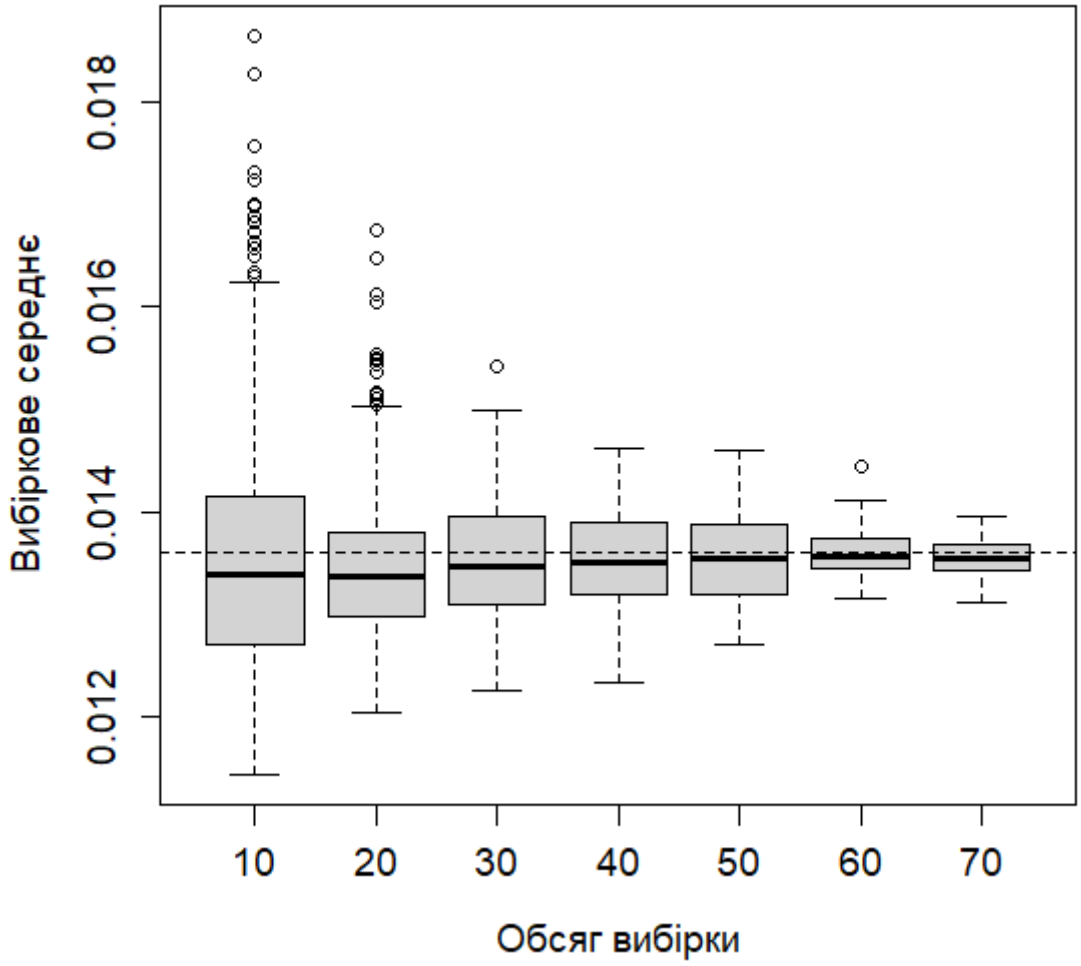

Рисунок 2.6 – Розподіл вибіркового середнього значень формальдегіду

З графіків видно, що малі обсяги вибірок недооцінюють значення вибіркового середнього, тому для отримання обґрунтованих значень показників доцільно використовувати вибірки обсягом більше ніж 50 елементів.

Рисунки 2.7 – 2.9 іллюструють вибіркові середні значення (суцільні лінії) і їх 95% довірчі інтервали, які є несиметричними і помітно звужуються при збільшенні розміру вибірки.

Таким чином, проведені дослідження показали, що малі вибірки недооцінюють значення середнього і призводять до ненадійних оцінок.

Проте у випадках, коли малий обсяг вибірки є неминучим доцільно використовувати бутстреп метод для обчислення параметрів описової

статистики та їх довірчих інтервалів. Також метод дозволяє проводити аналіз експериментальних даних зі складними законами розподілу.

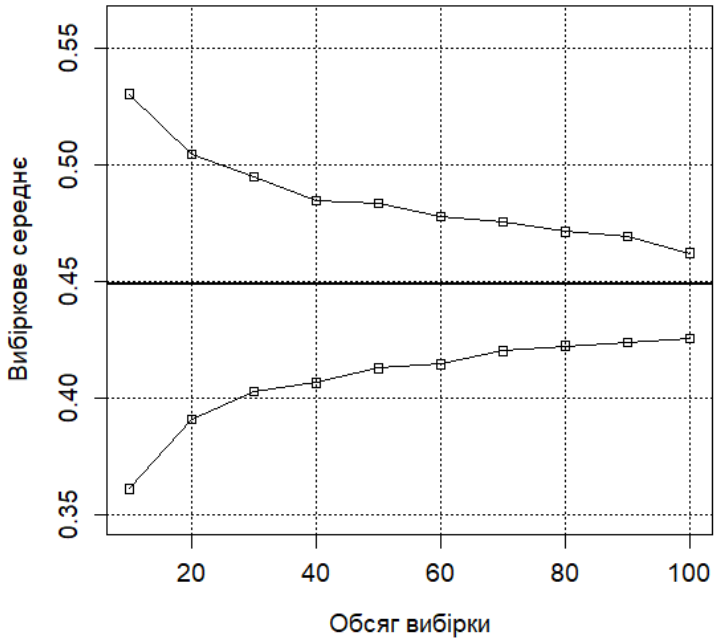

Рисунок 2.7 – Вибіркове середнє і його 95% довірчий інтервал для значень концентрації пилу

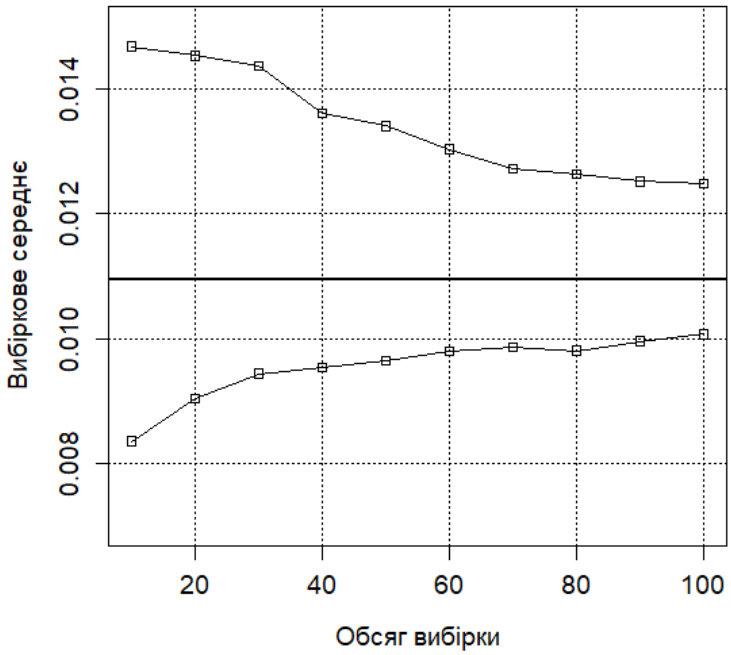

Рисунок 2.8 – Вибіркове середнє і його 95% довірчий інтервал для значень фенолу

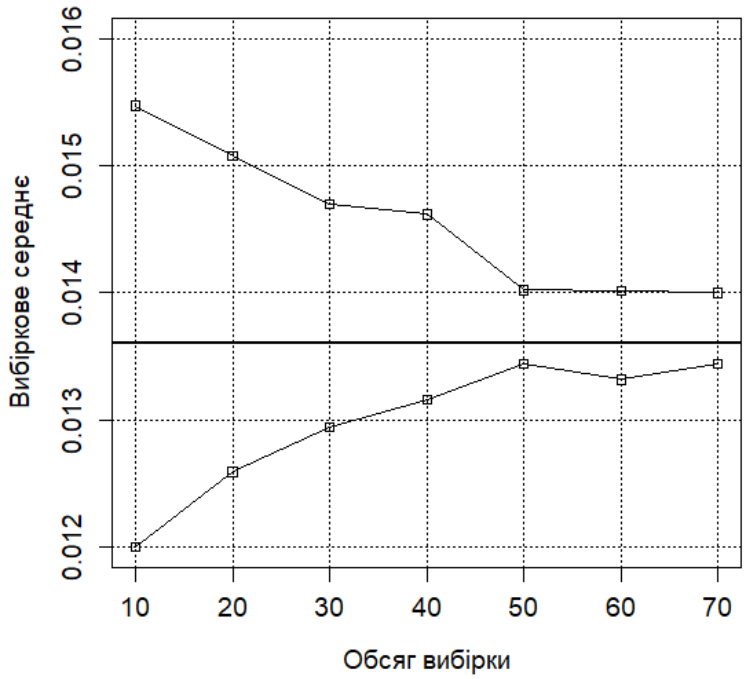

Рисунок 2.9 – Вибіркове середнє і його 95% довірчий інтервал для значень формальдегіду

Для реалізації бутстреп методу можуть бути рекомендовані мова програмування R.

ВИСНОВКИ

Бутстреп з'явився для того, щоб боротися зі зміщенням емпіричних даних. Потім з'ясувалося, що його варто використати для оцінки вибіркової дисперсії, довірчих інтервалів та перевірки гіпотез. Отже, це універсальний метод.

Бутстреп придатний до роботи з будь-якими статистичними завданнями, чи то перевірки гіпотез про закони розподілу випадкових величин, регресія, дисперсійний аналіз чи багатовимірна класифікація даних. Звичайно, у багатьох випадках у бутстрепі немає необхідності, та й обходиться він недешево, оскільки потребує великого обсягу обчислень. Нині побудова довірчих інтервалів є найбільш важливою практичною стороною використання цього методу.

Доведено, що технологія бутстреп має низку переваг перед класичною. Використання цього методу дає можливість багаторазового розмноження вибірки. Це, зокрема, дає змогу отримати досить надійні оцінки статистичних параметрів емпіричних даних.

ПЕРЕЛІК ПОСИЛАНЬ

1. Ермакова С. М. Метод Монте-Карло и смежные вопросы. Москва : Наука, 1974. 248 с.

2. Иванова И. М. Случайные числа и их применения. Москва : Финансы и статистика, 1984. 109 с.

3. Михайлов Г. А. Оптимизация весовых методов Монте-Карло // *Сибирский математический журнал*. 2004. Том 45. № 2.

4. Соболь И. М., Мышецкая Е. Е. Об использовании квази-Монте-Карло в оценках bootstrap // *Математическое моделирование*. 2004. № 16 (2). С. 118 – 122.

5. Шитиков В. К., Розенберг Г. С. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. Тольятти : Кассандра, 2013. 314 с.

6. Эфрон Б. Нетрадиционные методы многомерного статистического анализа: сборник статей. Москва : Финансы и статистика, 1988. 263 с.

7. Efron B. Bootstrep methods. Another look at the Jacknife // *The Annals of Statistics*. 1979. Vol. 7. № 1. P. 1 – 26.

8. Efron B., Gong G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation // *The American Statistician*. 1983. Vol. 37. № 1. P. 36 – 48.

9. Efron B. Better bootstrap confidence intervals // *Journal of the American Statistical Association*. 1979. № 82. P. 171 – 200.

10. Efron B., Tibshirani R. J. An introduction to the bootstrap. New York : Chapman and Hall, 1993. 436 p.

11. Manly B. F. J. Randomization, bootstrap and Monte Carlo methods in biology. London : Chapman & Hall, 2007. 445 p.

12. Varian H. Bootstrap Tutorial // *Mathematica Journal*. 2005. № 9. P. 768 – 775.

13. Weisstein Eric W. Bootstrap Methods. URL : http://mathworld. wolfram.com/BootstrapMethods.html.

14. Saigo H. Comparing Four Bootstrap Methods for Stratified Three-Stage Sampling // *Journal of Official Statistics*. 2010. Vol. 26. No. 1. P. 193 – 207.

15. Kirk P. Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data // *Bioinformatics*. 2009. № 25 (10). P. 1300 – 1306. URL : doi:10.1093/bioinformatics/btp139.

16. Никитенко А. Оценка методом бутстреп. URL : https://craftapp mobile.com/bootstrap-evaluation/.

ДОДАТОК А

Код програми

```
library(mgcv)
library (MASS)
library(gamlss)
library(pscl)
library(boot)
library(fitdistrplus)
library(rcompanion)
library(car)
library(DTK)
library(pid)
library(nortest)

md1 <-read.table("d5.txt", sep= "\t", head=TRUE)
md1<-data.frame(md1)
md1
md1 <-read.table("d6.txt", sep= "\t", head=TRUE)
lillie.test(md1$for.)
lillie.test(md1$fenol)

summary(md1)
summary(md1$for.)
hist(md1$for.,breaks = 230)
hist(md1$for.,breaks = 230, freq = F,col = "lightblue",
xlab = "Значення концентрації формальдегіду",
ylab = "Щільність ймовірностей")
```

```
 main = "Гістограма та
крива щільності ймовірностей")
lines(density(dust1,bw=0.4 ), col ="red", lwd = 2)
text(40,0.25,"a)",cex=1.3)
f1<-md1$for.
mean(f1)

outp10 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,10))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp10[i,] = c(i,mean(boot))
}
outp20 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,20))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp20[i,] = c(i,mean(boot))
}
outp30 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,30))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp30[i,] = c(i,mean(boot))
```

```
}
outp40 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,40))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp40[i,] = c(i,mean(boot))
}
outp50 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,50))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp50[i,] = c(i,mean(boot))
}
outp60 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,60))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp60[i,] = c(i,mean(boot))
}
outp70 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,70))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
```

```
mean(boot)
outp70[i,] = c(i,mean(boot))
}
outp80 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,80))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp80[i,] = c(i,mean(boot))
}
outp90 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,90))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp90[i,] = c(i,mean(boot))
}
outp100 = matrix(nrow=1000, ncol=2)
for(i in 1:1000){
v1<-c(sample(f1,100))
boot <-numeric(1000)
for (j in 1:1000) boot[j] <-(mean(sample(v1,replace=T)))
mean(boot)
outp100[i,] = c(i,mean(boot))
}
pff10 <-sort(outp10[,2])
pff20 <-sort(outp20[,2])
pff30 <-sort(outp30[,2])
```

```
pff40 <-sort(outp40[,2])
pff50 <-sort(outp50[,2])
pff60 <-sort(outp60[,2])
pff70 <-sort(outp70[,2])
pff80 <-sort(outp80[,2])
pff90 <-sort(outp90[,2])
pff100 <-sort(outp100[,2])
lp2<-cbind(pm10,pm20,pm30,pm40,pm50,pm60,pm70,pm100)
lp2

lp2<-as.data.frame(lp2)
names(lp2)<-c("10","20","30","40","50","60","70","")
lp1

par(fin=c(4,4))
boxplot(lp2,xlab="Обсяг вибірки",
 ylab="Вибіркове середнє",cex.axis=1.2, cex.lab=1.2)
abline (h=0.4492,lwd=2)
```