КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет комп'ютерних наук та кібернетики
Кафедра прикладної статистики

Кваліфікаційна робота
на здобуття ступеня бакалавра

за спеціальністю 124 Системний аналіз

на тему:
АНАЛІЗ НЕЙМОВІРНІСНИХ ВИБІРКОВИХ ОБСТЕЖЕНЬ

Виконав студент 4-го курсу
Карауш Марко Максимович

(підпис)

Науковий керівник:
доцент, доктор фіз.-мат. наук
Розора Ірина Василівна

(підпис)

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент

(підпис)

Роботу розглянуто й допущено до захисту на засіданні кафедри прикладної статистики
«5» червня 2023 р.,
протокол № 11
Завідувач кафедри
І. В. Розора

(підпис)

Київ – 2023

РЕФЕРАТ

Реферат до роботи "Аналіз неймовірнісних вибіркових обстежень" студента 4 курсу Карауша Марка Максимовича, наукового керівника Розори І.В.

Робота складається з 44 сторінок, містить 7 ілюстрацій. Використано 14 джерел. Додатків до роботи 5.

Ключові слова: НЕЙМОВІРНІСНІ ВИБІРКОВІ ОБСТЕЖЕННЯ, МЕТОД ТИПОВИХ ПРЕДСТАВНИКІВ, КВОТНА ВИБІРКА, ВИБІРКА "ГНІЗДОВА", МЕТОД "СНІГОВОЇ ГРУДИ", СТІХІЙНА ВИБІРКА, МЕТОД ПІДСТАНОВКИ, МЕТОД ЗВАЖЕНИХ ОЦІНОК, МЕТОД БУТСТРЕП.

Об'єкт дослідження: неймовірнісні вибіркові обстеження.

Мета роботи: дослідження теоретичних аспектів та використання неймовірнісних вибіркових обстежень.

Методи та інструменти дослідження: використання різних методів неймовірнісних вибіркових обстежень за допомогою мови R.

Результати та їх новизна: розглянуто та проаналізовано різні методи використання неймовірнісних вибіркових обстежень, зокрема для оцінки пропорцій , їх переваги та особливості.

Інформація щодо впровадження: результати роботи можуть бути використані в соціології, маркетингу, статистиці та інших галузях, де використовуються методи вибіркових обстежень.

Взаємозв’язок з іншими роботами: робота базується на сучасних методах неймовірнісних вибіркових обстежень.

Рекомендації щодо використання результатів роботи: результати роботи можуть бути використані для покращення процесу збору та аналізу даних в різних галузях, що використовують методи вибіркових обстежень.

Сфера застосування: соціологія, маркетинг, статистика.

Значимість роботи: робота важлива для сучасного світу, де інформація стала основним товаром і важливим ресурсом. Здатність збирати та аналізувати великі обсяги даних стала не просто потребою, а вирішальним фактором успіху.

Висновки та пропозиції щодо розвитку об’єкта дослідження (розроблення) та доцільності продовження досліджень або розробок: в майбутньому, такий підхід можна буде розглянути, наприклад, для написання Магістерської кваліфікаційної роботи.

ЗМІСТ

ВСТУП

У сучасному світі, де дані є центральним елементом прийняття рішень, важливість якісного та ефективного збору даних не можна переоцінити. В контексті статистичного аналізу, вибіркові обстеження є основним інструментом для збору даних, які використовуються для висновків про більшу популяцію. Загалом, вибіркові обстеження поділяються на дві основні категорії: імовірнісні та неімовірнісні.

У 1934 році з'явилася північна зірка теорії ймовірнісного дослідження. Єжи Нейман опублікував статтю в Королівському статистичному товаристві, в якій представив свій підхід до логічного висновку, оснований на дизайні (Neyman, 1934). Він зацікавив статистиків свого часу, що призвело до подальшого розвитку цієї теорії. Сучасні національні статистичні організації, наприклад, Статистика Канади і Французький Національний інститут статистики та економічних досліджень (INSEE), широко використовують ймовірнісні дослідження для отримання необхідних даних про цільову популяцію.

Узагальнена захопленість ймовірнісними дослідженнями у сфері офіційної статистики часто пов'язують з непараметричним характером методу висновку, який запропонував Нейман (1934). Тобто, ймовірнісні опитування дозволяють виводити обґрунтовані висновки про популяцію без залежності від модельних припущень. Цей аспект є особливо привабливим - а за словами Девіля (1991), навіть фундаментальним - для національних статистичних агентств, відповідальних за виробництво офіційної статистики. Важливо зазначити, що ці агентства традиційно уникають непотрібних ризиків, невід'ємних для підходів, що базуються на обґрунтованості припущень моделі, особливо коли основні припущення складно перевірити.

Оцінки, отримані в результаті ймовірнісних опитувань, можуть виявитися непродуктивними або навіть непридатними для використання, особливо при незначних розмірах вибірки (дивіться, наприклад, Rao та Molina, 2015). Більше того, вони засновані на припущенні, що помилки, не залежні від вибірки, такі як помилки вимірювання, охоплення або відсутності відповіді, мають незначний вплив.

У останні роки в національних статистичних агентствах почали п зміни, а інші джерела даних все більше привертають увагу. За цим трендом стоять п'ять основних факторів: i) спад відгуків на ймовірнісні опитування протягом останніх років; ii) висока вартість збору даних; iii) зростання навантаження на респондентів; iv) бажання мати доступ до статистики «в реальному часі» (Rao, 2020), тобто змогу створювати статистику майже одночасно або дуже швидко після виникнення інформаційних потреб; v) поширення неімовірнісних джерел даних (Rancourt, 2019), таких як адміністративні джерела, соціальні медіа, веб-опитування і так далі.

Деякі опитування, які проводять національні статистичні агентства, відзначаються дуже низькими показниками відповідей, і стає ризиковано покладатися лише на збір даних та методи оцінювання для компенсації потенційних упереджень, що виникають через відсутність відповідей. Дійсно, кілька авторів (наприклад, Rivers, 2007; Elliott та Valliant, 2017) звернули увагу на схожість між ймовірнісним опитуванням з екстремально низьким рівнем відповідей і неімовірнісним опитуванням. Проте, неімовірнісне опитування має свої переваги, оскільки зазвичай має значно більший розмір вибірки і є менш вартісним.

Ця дипломна робота зосереджується на аналізі неімовірнісних вибіркових обстежень. Зокрема, ми хочемо розкрити поняття неімовірнісних вибіркових обстежень, виявити їх переваги над

імовірнісними, а також на практиці дослідити використання неімовірнісних вибірок для аналізу пропорцій.

Мета: аналіз неімовірнісних вибіркових обстежень та практичне використання неімовірнісних вибірок для аналізу пропорцій.

Об’єкт : аналіз обстежень .

Суб’єкт: неімовірнісні вибірки

1 ТЕОРЕТИЧНІ АСПЕКТИ НЕЙМОВІРНИСНИХ ВИБІРКОВИХ ОБСТЕЖЕНЬ

Неймовірнісна (невипадкова) вибірка - це метод відбору одиниць вибіркової сукупності, принцип якого відрізняється від випадкового. Як і в разі з вірогідним відбором, основна мета невипадкового відбору полягає у отриманні сукупності, яка репрезентує об'єкт дослідження. Однак, на відміну від вірогідної вибірки, статистичні висновки про всю множину об'єктів в цьому випадку робити не зовсім вірно. Ці висновки так чи інакше вірні лише для генеральної сукупності, яка не завжди збігається з об'єктом дослідження.

Основні аспекти неймовірнісних вибіркових обстежень включають:

- **Відсутність випадковості**: Неймовірнісні вибірки не використовують випадковий відбір. Замість цього елементи вибірки обираються на основі їх доступності, зручності, або за іншими критеріями, що встановлює дослідник.
- **Сміщення вибірки**: Такі вибірки мають схильність до сміщення, оскільки не всі члени цільової популяції мають однакову можливість бути включеними в вибірку.
- **Відсутність можливості оцінки невизначеності**: При використанні неймовірнісних вибіркових обстежень дослідники не можуть обчислити статистичну невизначеність або встановити інтервали довіри для їх результатів, оскільки відсутність випадкового відбору перешкоджає коректному використанню статистичних методів.
- **Низька загальна відповідність**: Оскільки неймовірнісні вибірки не використовують випадковий відбір, вони можуть не точно представляти цільову популяцію, і тому результати можуть бути менш загальноприйнятними.

1.1 ФОРМИ НЕЙМОВІРНІСНОГО ВІДБОРУ

Виділяють два основних види невипадкової вибірки (відбору): **цілеспрямований** (відбір на власний розсуд) та **стихійний**.

Формами невірогідного або цілеспрямованого відбору виступають: метод типових представників, квотна вибірка, вибірка гніздування та метод "снігової кулі".

1.1.1 МЕТОД ТИПОВИХ ПРЕДСТАВНИКІВ

Метод типових представників, або типова вибірка (judgemental sampling), це детермінована методика збору даних, де "експерт" робить вибірку, орієнтуючись на власні висновки з метою досягнення репрезентативності.

Типова вибірка більше ґрунтується на особистій оцінці дослідника, ніж на випадковому виборі елементів. Дослідник може свідомо або навмисно вирішувати, які елементи включити до вибірки. Застосування типової вибірки дозволяє отримати докладну оцінку характеристик генеральної сукупності. Однак цей метод не дає можливості об'єктивно оцінити точність результатів дослідження. Оскільки не можна визначити ймовірність включення кожного окремого елемента до вибірки, отримані результати не можуть бути статистично поширені на всю сукупність.

Наприклад, покупці торгового центру можуть бути відображенням населення міста, або декілька міст можуть представляти країну.

В таких ситуаціях можуть бути різні спотворення, як очевидні, так і приховані. Наприклад, якщо використовується метод перехоплення у торговому центрі, то вибірка буде містити надмірну кількість людей, які часто відвідують магазин, або тих, у кого багато вільного часу. Більше того, не існує методу для кількісної оцінки таких спотворень, оскільки основа

вибірки невідома, а процедура її формування не чітко визначена. Тому зазвичай використовують інші методи, як-от квотна вибірка.

1.1.2 КВОТНА ВИБІРКА

Квотна вибірка (quota sampling) - це метод збору даних, що включає двоступінчасту обмежену вибірку. Перший крок полягає в утворенні контрольних груп, або квот, із елементів генеральної сукупності. На другому етапі вибір елементів базується на зручності відбору або думці дослідника.

Відбір заснований на квотах передбачає попереднє визначення, на основі цілей дослідження, чисельності груп респондентів, що відповідають певним вимогам (ознакам). Репрезентативність квотної вибірки збільшується прямо пропорційно ступеню стабільності значень тих характеристик, за якими встановлюються квоти. Наприклад, для цілей дослідження було вирішено, що в супермаркеті має бути опитано п'ятдесят чоловіків і п'ятдесят жінок. Інтерв'юер проводить опитування до того моменту, поки не набере встановлену квоту.

Виділяють два типи квотного відбору:

- апріорний відбір (виконується інтерв'юером відповідно до квотного плану на етапі збору первинної інформації);
- апостеріорний відбір (проводиться для коригування вибірки: наприклад, при вуличному опитуванні серед відповідачів часто є відхилення за деякими параметрами (вік, стать тощо). У такому випадку можна врахувати отримані результати, або можна провести вибірку з вибірки квотним методом).

Переваги квотного відбору:

- відсутність необхідності в повторних візитах;

- можливість досягти заданої точності результатів при меншому обсязі вибірки (хоча це спірне питання, не всі дослідники погоджуються з цим твердженням).

Обсяг квотних вибірок. Вимоги до вибірки можуть бути жорсткими та зниженими. Жорсткі вимоги передбачають повне співпадіння пропорцій генеральної і вибіркової сукупностей за комбінаціями ознак. У цьому випадку структура вибіркової і генеральної сукупностей за вказаними параметрами точно співпадає. При використанні знижених (часткових) вимог контролюють лише співпадіння пропорцій за кожним параметром окремо.

Навіть якщо в структурі вибірки повністю відображена структура популяції з урахуванням контрольних характеристик, немає гарантії, що ця вибірка репрезентативна. Якщо характеристика, що безпосередньо пов'язана з проблемою дослідження і не врахована, то квотна вибірка нерепрезентативна. Важливі контрольні характеристики часто забуваються через те, що на практиці дуже важко включити велику кількість таких характеристик у вибірку. Елементи вибираються з кожної квоти, виходячи з зручності або на основі думки дослідника. Отже, існує велика ймовірність необ'єктивності при відборі. Інтерв'юери можуть поїхати в ті з вказаних районів, де найлегше знайти підходящих респондентів. Більше того, вони можуть уникати людей, які виглядають недружелюбно, погано одягнені або живуть у місцях, куди неудобно дістатися. Квотна вибірка не дозволяє оцінити величину помилки вибірки.

Застосовуючи вибірку за квотами, дослідник намагається отримати репрезентативну вибірку при порівняно низькому рівні витрат. Переваги такої вибірки - її низька вартість та зручність вибору елементів для кожної квоти. Останнім часом було введено більш суворий контроль за діями інтерв'юерів та процедурами проведення опитування, що дозволяє

зменшити спотворення при відборі. Запропоновані заходи щодо покращення якості вибірок за квотами при проведенні інтерв'ю у торгових центрах. За певних умов застосування вибірки за квотами дає результати, схожі на результати застосування випадкової вибірки.

Іноді у дослідників виникає спокуса збільшити число контрольованих квотних параметрів у надії на підвищення ступеня достовірності результатів. Однак на практиці це призводить до зростання систематичної помилки та ускладнює полеву роботу інтерв'юера.

1.1.3 ВИБІРКА "ГНІЗДОВА"

Вибірка "гніздова" (або кластерна, серійна) передбачає відбір не окремих об'єктів дослідження (наприклад, людей), а цілих груп. Ці групи вибираються випадковим чином, а потім усередині цих груп проводиться загальне опитування. Наприклад, у вищому навчальному закладі з великою кількістю студентських груп вибірку можна здійснювати шляхом випадкового відбору цих груп і подальшого загального опитування у цих групах.

Понятно, що інтервал достовірності при гніздовій вибірці буде меншим (вибірка точніша) при тій самій надійності, ніж при випадковій, оскільки міжгрупова дисперсія менше загальної дисперсії. Внутрішньогрупова дисперсія нас не цікавить, оскільки ми опитуємо все гніздо цілком, і тому не маємо відхилень вибіркового показника від генерального всередині цієї групи. Отже, нас повинно хвилювати лише те, наскільки правильно ми вибрали самі групи. Саме тому ми беремо до уваги лише міжгрупову дисперсію.

Розглянемо приклад створення списку для опитування 1000 осіб (розмір вибірки) для вивчення громадської думки населення міста Київ. Не

маючи списку всіх жителів міста, можна почати з отримання карти міста, щоб визначити всі його квартали і скласти їх список. Список кварталів стає основою вибірки, з якої випадковим чином або систематично здійснюється відбір кварталів. Потім планується вибірка житлових будинків з кожного кварталу, встановлюється зв'язок з сім'ями, що проживають у відібраних будинках, і у представника кожної сім'ї проводиться інтерв'ю для опитувального листа. Припустимо, є 500 кварталів; з них випадковим чином відібрано 25. У цих кварталах ідентифіковано 4000 сімей. Зв'язок встановлюється з 25 відсотками цих сімей, оскільки потрібна вибірка з 1000 осіб, тобто (4000 сімей помножити на 0,25 дорівнює 1000). Ці представники 1000 сімей відбираються випадковим або систематичним чином.

1.1.4 МЕТОД "СНІГОВОЇ ГРУДИ"

Метод **"снігової груди".** Цей метод застосовується для відбору експертів та рідко зустрічаючихся груп респондентів ("рідкісних елементів"), наприклад, споживачів з високим рівнем доходу, або представників елітних груп. По суті, це техніка пошуку та відбору респондентів з певним набором властивостей у таких умовах, коли важко визначити межі генеральної сукупності. Особливість методу полягає в тому, що крім першого етапу, вибір кожного наступного респондента відбувається за рекомендацією респондентів, включених у склад вибірки на попередньому етапі. Кожен з респондентів показує інтерв'юеру, де можна знайти людей, які його цікавлять (іноді навіть сам зв'язується з ними і рекомендує інтерв'юера), і вибірка з кожним етапом розростається накшталт снігової груди, тобто - це метод, за яким випадковим чином відбирається початкова група респондентів. В подальшому відбір здійснюється серед кандидатів, вказаних першими респондентами, або на основі наданої ними інформації. Цей процес проходить хвильовим чином,

коли респонденти, що пройшли опитування, називають наступних кандидатів і т.д.

Метод "снігової груди" можна відобразити на такому прикладі. Припустимо, ви досліджуєте групу редкісних колекціонерів монет у великому місті. Вам важко визначити всю генеральну сукупність, бо ви не знаєте, хто з городян є колекціонерами.

Ви починаєте з кількох осіб, кого ви випадковим чином визначили як колекціонерів. Ви проводите з ними інтерв'ю і просите їх рекомендувати інших колекціонерів, яких вони знають. Вони надають вам контакти кількох своїх знайомих, і ви розширюєте свою вибірку, зв'язуючись з цими новими особами. Таким чином, ваша вибірка поступово збільшується, накшталт снігової груди, що котиться з гори.

Цей процес повторюється доти, доки ви не досягнете потрібного розміру вибірки - 1000 респондентів. Кожен наступний респондент вибирається на основі рекомендацій попередніх, що дозволяє ефективно виявляти і залучати до дослідження "рідкісні елементи" генеральної сукупності.

1.1.5 СТІХІЙНА ВИБІРКА

Стіхійна вибірка, відома також як неформальна вибірка або вибірка зручності, відбувається, коли дослідник вибирає елементи для включення до вибірки на основі їх доступності або легкості доступу. Це може бути корисним, коли дослідникам треба швидко зібрати дані, але цей метод може привести до зміщення вибірки, оскільки вибірка може не відображати належним чином генеральну сукупність.

Розглянемо приклад. Припустимо, ви - студент, який проводить дослідження про уподобання в їжі серед студентів вашого університету. Ви вирішуєте швидко провести опитування, запитуючи у своїх товаришів по

гуртожитку, що вони найчастіше їдять на сніданок. Це стихійна вибірка, оскільки ви вибрали опитувати людей, які вам зручно опитувати.

Проте, цей метод має значні обмеження. Наприклад, ваша вибірка може бути зміщеною, оскільки студенти з гуртожитків можуть мати інші харчові звички порівняно з тими, хто живе поза кампусом. Таким чином, хоча цей метод може бути зручним, він не завжди надає точні або репрезентативні дані.

,

1.2 МЕТОДИ Й ТЕХНІКИ ЗБОРУ ДАНИХ ДЛЯ НЕЙМОВІРНИСНИХ ВИБІРКОВИХ ОБСТЕЖЕНЬ

Для збору даних у неймовірнісних вибіркових обстеженнях можуть використовуватися різні методи та техніки, в залежності від конкретних цілей дослідження і доступності респондентів. Ось деякі з них:

1. **Опитування та анкетування**: Ці методи дозволяють швидко і ефективно зібрати інформацію від великої кількості респондентів. Вони можуть бути дуже структурованими, що дає змогу аналізувати дані кількісно, але в той же час можуть обмежувати відповіді, тому що респонденти відповідають лише на певні запитання.

2. **Інтерв'ю**: Інтерв'ю дає можливість отримати більш глибоку інформацію та вивчити думки і погляди людей детальніше, ніж через опитування. Проте вони потребують більше часу і ресурсів для проведення та аналізу.

3. **Обстеження**: Обстеження можуть бути використані для вивчення фізичного середовища або соціального контексту. Вони можуть бути корисними для визначення умов життя, доступності ресурсів або рівня соціального розвитку.

4. **Аналіз документів або записів**: Цей метод дозволяє вивчити історичний контекст, соціальні та культурні тенденції, а також отримати доступ до інформації, яка може бути недоступна через інші методи.

5. **Спостереження**: Спостереження дозволяє досліднику вивчити поведінку людей в їх природному середовищі. Це може бути корисно для вивчення взаємодій, поведінкових шаблонів або соціальних процесів.

6. **Експерименти**: Експерименти дозволяють тестувати гіпотези в контрольованих умовах. Вони можуть бути використані для визначення причинно-наслідкових зв'язків, але вони мають обмеження у відображенні реального життя.

7. **Фокус-групи**: Група людей запрошується обговорити певну тему під керівництвом модератора.

Кожен з цих методів має свої переваги та недоліки, і вибір конкретного методу буде залежати від цілей дослідження, доступності респондентів та ресурсів дослідника.

1.3 ОСОБЛИВОСТІ Й ПЕРЕВАГИ НЕЙМОВІРНИСНИХ ВИБІРКОВИХ ОБСТЕЖЕНЬ

Неймовірнісні вибіркові обстеження мають ряд особливостей і переваг, але вони також мають свої недоліки, особливо порівняно з ймовірнісними вибірками.

Переваги неймовірнісних вибірок включають:

- Легкість в зборі: Неймовірнісні вибірки, зазвичай, легше зібрати, оскільки вони не вимагають складних механізмів вибірки.
- Швидкість: Вони часто дозволяють швидко зібрати дані, особливо коли час - критичний фактор.

- Низька вартість: Вони можуть бути дешевшими, ніж ймовірнісні вибірки, оскільки не потребують великих витрат на планування та виконання вибірки.

Недоліки неймовірнісних вибірок включають:

- Сміщення вибірки: Неймовірнісні вибірки часто мають сміщення вибірки, оскільки не всі члени популяції мають однакову ймовірність бути вибраними.
- Обмеження узагальнення: Результати неймовірнісних вибірок часто не можуть бути узагальнені на більшу популяцію.
- Відсутність міри невизначеності: Неймовірнісні вибірки не дозволяють розрахувати інтервали довіри або виміряти статистичну невизначеність.

Приклад:

Дослідник хоче вивчити зв'язок між зайнятістю та станом здоров'я в місті. Якщо він використовує неймовірнісну вибірку, він може просто підійти до людей на вулиці і запитати їх про їх зайнятість та стан здоров'я. Це швидко і легко, але може призвести до сміщення вибірки, оскільки люди, які мають роботу, можуть бути менш доступні для інтерв'ю протягом дня.

З іншого боку, якщо дослідник використовує ймовірнісну вибірку, він має можливість узагальнити свої результати на більшу популяцію. Але це зазвичай вимагає більше часу, більше витрат на планування та реалізацію вибірки, а також вибірку має бути відносно великою, щоб забезпечити достатній рівень точності.

Зрештою, обраний дослідником метод вибірки буде залежати від його конкретних цілей та ресурсів. Якщо головна ціль - швидко зібрати дані з невеликою витратою ресурсів, то неймовірнісна вибірка може бути найкращим вибором. Проте, якщо головна ціль - забезпечити можливість

узагальнення результатів на більшу популяцію з мірою невизначеності, то ймовірнісна вибірка буде більш придатною.

2 ВИКОРИСТАННЯ НЕЙМОВІРНІСНИХ ВИБІРОК ТА ДАНИХ ДЛЯ ОЦІНКИ ПРОПОРЦІЙ

Нерепрезентативні вибірки часто використовуються в ситуаціях, коли неможливо отримати репрезентативну вибірку або коли така вибірка не є необхідною для досягнення цілей дослідження. Нижче наведено кілька прикладів, коли можуть використовуватися нерепрезентативні вибірки:

- Дослідження пілотажного характеру: Наприклад, перед тим, як розгорнути велике масштабне дослідження, дослідник може виконати невеличке попереднє дослідження з використанням нерепрезентативної вибірки, щоб визначити, чи варто продовжувати більш велике дослідження.
- Швидкі або термінові дослідження: Якщо досліднику потрібні швидкі або негайні результати, він може використати нерепрезентативну вибірку. Наприклад, в ситуаціях кризи або невідкладних ситуацій.
- Дослідження з високою ступенем специфікації: Якщо дослідження має дуже специфічну ціль, наприклад, вивчення певного вузького сегменту населення, то може бути використана нерепрезентативна вибірка.
- Експлоративні дослідження: Якщо мета дослідження - це попереднє вивчення явища, тоді нерепрезентативна вибірка може бути достатньою.

Переваги використання нерепрезентативних вибірок для оцінки пропорцій включають:

- Швидкість: Нерепрезентативні вибірки зазвичай швидше збираються, ніж репрезентативні вибірки.
- Економія ресурсів: Збір нерепрезентативних вибірок зазвичай вимагає менше часу, грошей та інших ресурсів.
- Доступність: Нерепрезентативні вибірки можуть бути єдиним доступним варіантом в ситуаціях, коли доступ до генеральної сукупності обмежений або неможливий.
- Проведення попередніх досліджень: Нерепрезентативні вибірки можуть бути корисними для проведення попередніх досліджень, щоб визначити, чи варто проводити більше часу та ресурсів на повномасштабне дослідження.

З іншого боку, важливо зауважити, що нерепрезентативні вибірки мають свої обмеження. Вони можуть не дати точного представлення про генеральну сукупність і можуть бути схильними. Якщо ці обмеження прийнятні для конкретного дослідження, використання нерепрезентативних вибірок може бути ефективною стратегією.

2.1 МЕТОДИ ОЦІНКИ ПРОПОРЦІЙ

Оцінка пропорцій з використанням нерепрезентативних вибірок та даних може бути досягнута за допомогою різних методів. Деякі з найпоширеніших методів включають наступне:

2.1.1 МЕТОД ПІДСТАНОВКИ

Метод підстановки (**Imputation Method**) зазвичай використовується в контексті обробки відсутніх даних. Його суть полягає в тому, щоб заповнити відсутні значення в наборі даних на основі інших доступних даних. Є різні стратегії підстановки, включаючи просте заповнення

середнім значенням, підстановку на основі моделі (де відсутні дані заповнюються на основі прогнозування моделі, що вивчена на основі наявних даних), або множинну підстановку, де створюються кілька версій набору даних, кожна з різними підстановками, і результати аналізу згодом усереднюються.

2.1.1.1 СЕРЕДНЄ ЗНАЧЕННЯ

Один із підходів цього методу — це "**Mean Imputation**", де пропущені значення замінюються середнім значенням відповідної змінної.

Отже, базова формула для цього методу досить проста:

X_filled = X_missing

де **X_filled** - заповнене значення, **X_missing** - пропущене значення.

Припустимо, що у нас є набір даних з віком та статтю людей і ми хочемо знайти вікову пропорцію, але деякі значення віку відсутні. Ми можемо використати метод підстановки для заповнення цих пропусків.

Для цього спочатку встановлюємо та завантажуємо пакет mice . Наводимо набір даних з двома стовбцями "**age**" та "**gender**" , з пропущеними значеннями позначеними як **NA**. Далі за допомогою методу mean пропущені значення віку замінюються на середнє значення віку для всіх існуючих записів та для порівняння результатів виводимо оригінальний та фінальний набори даних. Залишається тільки зробити вікові групи за допомогою cut ("**Young**" для віку 0-30, "**Middle**" для віку 30-40, та "**Old**" для віку більше 40) та розрахувати пропорцію кожної групи відносно загальної кількості спостережень.

Вивід :

```
> # Виводимо оригінальні та заповнені дані
> print(data)
   age gender
1   25      M
2   27      F
3   NA      M
4   30      F
5   32      M
6   35      F
7   NA      M
8   40      F
9   43      M
10  47      F
11  NA      M
> print(complete(imputed_data))
      age gender
1  25.000      M
2  27.000      F
3  34.875      M
4  30.000      F
5  32.000      M
6  35.000      F
7  34.875      M
8  40.000      F
9  43.000      M
10 47.000      F
11 34.875      M
```

Пропорція:

```
> print(age_proportion)
age_groups
    Young    Middle       Old
0.2727273 0.5454545 0.1818182
```

2.1.1.2 ЛІНІЙНА РЕГРЕСІЯ

Зазвичай процес оцінки передбачує, що спершу модель регресії встановлюється на базі спостережуваних даних, а потім, за допомогою вагових коефіцієнтів регресії, відбувається прогнозування та заміщення відсутніх значень.

У цілому, регресійний метод заміни відсутніх значень в таблицях даних включає наступні елементи. Позначимо дані як **Y**= (y_1,y_2,...,y_n) - це вектор результуючих значень вивченої змінної, **X**= (x_1, x_2, ... x_k) - це вектор пояснювальних змінних (предикторів), пов'язаних з **Y**. Через **θ** позначаємо невидимі векторні величини або параметри досліджуваної сукупності, а через $\tilde{\mathbf{y}}$ - невідомі, але потенційно спостережувані значення змінної.

Найпростіший і найбільш розповсюджений варіант цієї моделі є проста множинна лінійна модель, в якій розподіл **Y** для заданого **X** є нормальним із середнім, що є лінійною функцією X:

$$E(y_i|\beta,X) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} \text{ для i} = 1,...,n.$$

В даній ситуації **Y** представляє безперервні величини, змінні **X** можуть бути дискретними або безперервними, а **θ** = (β_0, β_1,..., β_k) . Для визначення коефіцієнтів **β** використовують тільки повні або комплектні (без пропусків) дані за методом найменших квадратів (МНК).

Знайдені коефіцієнти потім використовуються для відновлення пропущених значень. Гельман відзначає, що проблема полягає в тому, що обчислені дані не включають у себе член похибки, що входить до їх оцінки, тому ці оцінки досконало відповідають лінії регресії без будь-якої залишкової дисперсії. Це веде до зменшення варіативності (дисперсії) значень, зміцнення зв'язків (кореляції) між характеристиками і передбачає більшу точність замінених значень, ніж це об'єктивно виправдано.

Для цього прикладу замінюємо стовпець "**gender**" на "**income**", який відображає заробіток за рік . Створюємо копію даних та застосовуємо лінійну регресію для прогнозування пропущених значень "**age**" у основі "**income**". Обчислюємо та замінюємо пропущені значення та для порівняння виводимо початковий та фінальний набір .

Цей підхід використовує точний прогноз, заснований на моделі, і не враховує випадкові відхилення.

Вивід :

```
> # Виводимо оригінальні та заповнені дані
> print(data)
   age income
1   25  30000
2   27  35000
3   NA  40000
4   30  45000
5   32  50000
6   35  55000
7   NA  60000
8   40  65000
9   43  70000
10  47  75000
11  NA  80000
> print(data_imputed)
        age income
1  25.00000  30000
2  27.00000  35000
3  28.59898  40000
4  30.00000  45000
5  32.00000  50000
6  35.00000  55000
7  38.16244  60000
8  40.00000  65000
9  43.00000  70000
10 47.00000  75000
11 47.72589  80000
```

Пропорція :

```
> print(age_proportion)
age_groups
    Young    Middle       Old
0.3636364 0.3636364 0.2727273
```

2.1.1.3 МНОЖИННА ПІДСТАНОВКА

Множинна імпутація (**Multiple Imputation**) є розширенням одиничної імпутації, де відсутні дані заповнюються багато разів, щоб створити "множину" повних наборів даних. Процедура включає в себе три основні кроки:

- **Імпутація**: Відсутні дані заповнюються m разів, щоб створити m повних наборів даних.
- **Аналіз**: Кожен з m наборів даних аналізується окремо.
- **Комбінація**: Результати з m аналізів комбінуються за допомогою правил Рубіна.

Нехай δ — це параметр, для якого ми бажаємо знайти оцінку за допомогою аналізу. Враховуючи M отриманих даних, M оцінок δ:

$$(\hat{\delta}_1, \hat{\delta}_2, \dots, \hat{\delta}_m)$$

можна згенерувати та використати для обчислення наступних параметрів:

- Загальна оцінка – це середнє із індивідуальних точкових оцінок:

$$\hat{Q} = \frac{1}{M} \sum_{m=1}^{M} \hat{\delta}_m$$

- Дисперсія в межах групи для вибірки даних – це середнє із індивідуальних дисперсій:

$$U = \frac{1}{M} \sum_{m=1}^{M} Var(\hat{\delta}_m)$$

- Міжгрупова дисперсія між вибірками даних – це дисперсія оцінок:

$$B = Var(\hat{\delta}_1, \hat{\delta}_2, \dots, \hat{\delta}_m)$$

- Загальна дисперсія комбінованої оцінки, що включає як внутрішньогрупову, так і міжгрупову дисперсії:

$$T = U + \left(1 + \frac{1}{M}\right) B$$

- 95% інтервали для параметрів обчислюються як:

$$\hat{Q} \pm 1.96 * \sqrt{T}$$

Отже, Рубін вводить універсальний механізм для генерації множинних імпутацій пропущених значень, заснований на параметрах моделі генерації даних. В процесі такого підходу кожне відсутнє значення замінюється на декілька потенційних варіантів, утворюючи декілька наборів даних. Багаторазова імпутація передбачає заміну кожного пропущеного значення набором можливих значень. На початковому етапі датасет з пропущеними значеннями (або неповний датасет) реплікується кілька разів. Потім, на наступному кроці, відсутні значення замінюються обчисленими в кожному дублікаті датасету. Зауважте, що ці обчислені значення будуть варіюватися в різних копіях через випадкові відхилення (рахуючи з використанням генератора випадкових чисел). Це веде до створення множини датасетів. На останньому кроці, кожен отриманий набір даних аналізується окремо за допомогою стандартних методів, а результати об'єднуються для отримання неперекошених оцінок за допомогою правила Рубіна.

Питання про оптимальну кількість генерованих наборів даних є важливим. Стєф ван Бюрен стверджує, що кількість ітерацій може залежати від ступеня кореляції між змінними і відсотка пропущених даних в змінних. Він запропонував, що для досягнення надійної консистентності даних, зазвичай, достатньо провести від 5 до 20 ітерацій.

У цьому прикладі ми створюємо 5 різних наборів даних з імпутованими значеннями. Для цього створюємо 2 набори даних зі стовпцями:

- **success** - це бінарна змінна, яка відображає результат (наприклад, успіх або невдача) для кожного спостереження в датасеті. Це може бути будь-що - успішні продажі, виконані завдання, вдалі експерименти тощо. В деяких випадках це можуть бути пропущені значення, які ми хочемо імпутувати.
- **group** - це факторна змінна, яка розділяє дані на різні групи. Це може бути будь-яка змінна, яка розділяє дані на категорії, такі як група лікування у клінічному дослідженні, категорія продукту в продажах або вікова група в демографічному дослідженні.

Далі використовуємо множинну імпутацію за допомогою методу "pmm" для заміни пропущених значень . Створюємо функцію для розрахунку пропорції успіху в кожній групі . Наостанок розраховуємо пропорції успіху для кожного набору даних , отриманих в результаті імпутації та середню пропорцію успіху разом із стандартною помилкою.

Вивід :

```
> # Застосовуємо множинну імпутацію
> imp_data <- mice(data, m = 5, method = "pmm")

 iter imp variable
  1   1  success
  1   2  success
  1   3  success
  1   4  success
  1   5  success
  2   1  success
  2   2  success
  2   3  success
  2   4  success
  2   5  success
  3   1  success
  3   2  success
  3   3  success
  3   4  success
  3   5  success
  4   1  success
  4   2  success
  4   3  success
  4   4  success
  4   5  success
  5   1  success
  5   2  success
  5   3  success
  5   4  success
  5   5  success
```

Пропорція:

```
> # Виводимо результати
> print(data.frame(Proportion = props_mean, SE = props_se))
  Proportion        SE
1       0.58 0.0359011
```

2.1.2 МЕТОД ЗВАЖЕНИХ ОЦІНОК

Метод зважених оцінок - це узагальнення методу найменших квадратів, і його розвиток був поступовим процесом, до якого внесли свій вклад багато вчених.

Одним з важливих імен у цьому контексті є Карл Фрідріх Гаусс. Гаусс, відомий німецький математик та фізик, вніс величезний вклад в багато галузей науки. Він розвинув метод найменших квадратів для оцінки параметрів лінійних моделей, що є основою для методу зважених оцінок.

Сам Гаусс не ввів поняття "**зважених оцінок**", але його робота з методом найменших квадратів поклала фундамент для їхнього подальшого розвитку. Метод зважених оцінок використовує ту саму основну ідею, що й метод найменших квадратів - мінімізацію суми квадратів помилок - але з додаванням "**ваг**", які враховують різну надійність або точність окремих спостережень.

Практичні приклади застосування:

- Соціологічні опитування: Опитування можуть використовувати зважені оцінки для того, щоб забезпечити, що вибірка є репрезентативною щодо певних характеристик, таких як вік, стать, раса тощо.
- Економетрика: Зважені найменші квадрати є типовим методом для оцінки параметрів лінійних регресій, коли різні спостереження мають різну варіабельність.
- Медичні дослідження: В медичних дослідженнях зважена оцінка може бути використана для аналізу даних з когортних досліджень або клінічних випробувань, де різні учасники мають різні періоди спостереження.

Переваги:

- Зважена оцінка може бути більш точною, якщо ваги правильно відображають важливість спостережень.
- Вона може бути корисною, якщо деякі спостереження є більш релевантними або надійними для аналізу.

Недоліки:

- Якщо ваги встановлено неправильно, зважена оцінка може бути спотвореною.
- При наявності великої кількості даних визначення відповідних ваг може бути викликом.
- У певних випадках ваги можуть бути занадто суб'єктивними, що може призвести до упередженості в аналізі.

Наприклад, можемо взяти опитування проведене у місті Бостон у 2018 році. Було опитано 10000 людей на предмет наявності в них приватного страхування. У результаті маємо датасет з 4 стовпцями даних :

- **Age** – вік від 20 до 60 ;
- **Gender** – "F" чи "M" ;
- **Response** – відповідь , "Yes" чи "No" ;
- **Income** – дохід .

Припустимо , що ми хочемо дізнатись пропорцію , у кого є таке страхування , вважаючи , що для нас важливіші відповіді старших людей, тому хочемо надати більшу вагу їх відповідям .

Обчислюємо вагу для кожного спостереження на основі значень віку шляхом поділом значення **Age** кожного спостереження на суму всіх значень **Age** . Перевіряємо правильність розрахунків, здійснюючи сумування всіх значень ваг в датафреймі. Якщо ваги були розраховані правильно, то сума ваг має дорівнювати приблизно 1, оскільки ваги представляють відносну частку кожного спостереження в загальній сумі значень віку.

Після перевірки , отримали значення , тому вважаємо , що ваги встановлено вірно . Тепер, коли у нас є ваги для кожного спостереження, ми можемо розрахувати зважені пропорції відповідей "Так" для кожної групи за статтю:

У виводі отримали таблицю , яка надає пропорцію відповідей “Так” людей різної статі , на основі якої можна вирішувати поставлене питання .

```
# A tibble: 2 × 2
  Gender prop_yes
  <chr>     <dbl>
1 F         0.505
2 M         0.507
```

Але , як ми розуміємо , що дуже рідко вага буде визначатись тільки однією змінною , тому додаємо до нашого розрахунку змінну **Income** (дохід). Тобто робимо , щоб вага також враховувала дохід людини разом з віком .. А далі обчислюємо нову вагу шляхом множення **Age** на **Income** для кожного спостереження , а потім діленням отриманого добутку на суму добутків **Age** та **Income** для всього датафрейму . Перевіряємо правильність розрахунку та обчислюємо нову зважену пропорцію

У виводі отримали таблицю , яка надає пропорцію відповідей “Так” з вагою відносно віку та доходу .

```
# A tibble: 2 × 2
  Gender prop_yes
  <chr>     <dbl>
1 F         0.508
2 M         0.513
```

Отже , цей метод допомагає прийняти те чи інше рішення , враховуючи різну ‘вартість’ різних даних .

2.1.3 МЕТОД БУТСТРЕП

Метод бутстрепу - це підхід до статистичного висновку, що заснований на побудові великої кількості вибірок з вихідної вибірки з поверненням. Його основна мета - оцінити розподіл статистики вибірки та її варіативність без залучення строгих припущень про форму розподілу популяції.

Метод бутстрепу корисний в ситуаціях, коли класичні статистичні методи або непрактичні, або не можуть бути застосовані через недостатню

кількість даних, складність моделі або недостатні припущення про розподіл даних.

Переваги методу бутстрепу

- Він не вимагає жодних припущень про форму розподілу популяції.
- Його можна використовувати з малими вибірками для оцінки розподілу статистик.
- Він може бути застосований до великої кількості статистик, включаючи ті, для яких теоретичні розподіли важко обчислити.

Нехай ми маємо набір даних про вагу 50 людей. Ми хочемо оцінити медіану цих ваг, але на відміну від середнього, медіана може бути важкішою для аналізу, оскільки ми не можемо просто додати ваги разом і поділити на кількість людей.

За допомогою методу бутстрепу, ми можемо вибрати вибірки з вихідних даних (з поверненням), обчислити медіану для кожної вибірки, а потім дивитись на розподіл цих медіан, щоб отримати уявлення про те, як медіана може варіюватись в популяції.

Недоліки методу бутстрепу

- Для великих наборів даних бутстреп може бути обчислювально вимогливим, оскільки він вимагає генерації великої кількості вибірок.
- Метод бутстрепу може не працювати добре, якщо вибірка має сильну залежність між спостереженнями.
- Бутстреп може дати неточні оцінки на краях розподілу (так звані "хвости"), оскільки він маловірогідно, що відтворить рідкісні події з оригінальної вибірки.

Припустимо, ми проводимо соціологічне дослідження і хочемо вивчити доходи людей в регіоні. Відомо, що доходи часто мають довгий правий "**хвіст**" - є декілька людей з високими доходами, які значно перевищують середній рівень.

Якщо ми використаємо бутстреп, щоб оцінити розподіл доходів, то, взявши вибірки з поверненням з нашої вихідної вибірки, ми можемо отримати розподіл, який не враховує цих дуже високих доходів. Це відбувається тому, що шанси на те, що ці дуже високі доходи потраплять в нашу вибірку, не є великими, оскільки вони дуже рідкісні в порівнянні з більш звичайними доходами. Таким чином, наша оцінка розподілу доходів може бути неточною, оскільки вона не враховує ці "хвости".

Наприклад, беремо "Titanic dataset" з сайту Kagle . Цей датасет містить дані про пасажирів "Титаніку", такі як номер , значення виживання (1 чи 0) , стать , ПІБ, вік , та інше . Його часто використовують у навчальних чи тренувальних цілях.

Ми хочемо оцінити пропорцію вижившихжінок віком , які подорожували третім класом. Наша найкраща оцінка пропорції виживих у популяції - це 50%, але яка міра невизначеності в цій оцінці?

Ми виконаємо бутстреп, щоб оцінити довірчий інтервал цієї пропорції .

Для цього будемо використовувати бібліотеку "boot". Визначаємо нашу вибірку та створюємо функцію , яка обчислює пропорцію вижившихх.

results <- boot(data = women_third_class, statistic = boot_func, R = 5000): Цей рядок виконує бутстреп . Функція **boot** отримує вхідні дані **women_third_class**, статистику **boot_func** і кількість перезапусків (реплікацій) **R** (у даному випадку 5000). Вона генерує випадкові вибірки з **women_third_class**, обчислює статистику **boot_func** для кожної вибірки і повертає результати.

Вивід 1

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = women_third_class, statistic = boot_func, R = 5000)


Bootstrap Statistics :
    original        bias    std. error
t1*      0.5 -0.0003722222  0.04187356
```

Bootstrap Statistics: Цей розділ показує статистику бутстрепу для оцінки пропорції вижившх у вибірці **women_third_class**.

- **"Original"**: Це оригінальна статистика, яка була розрахована на основі оригінального набору даних. У прикладі, це пропорція вижившх жінок в 3 класі, яка складає 0.5.
- **"Bias"**: Це усереднена різниця між бутстреп-вибіркою і оригінальною статистикою. У нашому випадку, це дуже невелике значення, що означає, що наша бутстреп-вибірка має незначне зміщення від оригінального значення.
- **"Std. Error"**: Це стандартна помилка бутстреп-вибірки. Це показує, наскільки відмінні результати могли б отримати, якби виконували бутстреп-процедуру багато разів. У цьому випадку, стандартна помилка складає приблизно 0.0418.

Таким чином, ми можемо сказати, що пропорція вижившх жінок в третьому класі була близькою до 0.5 з невеликим зміщенням і стандартною помилкою близько 0.0418.

Вивід 2

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.4167,  0.5833 )
Calculations and Intervals on Original Scale
```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS: Цей розділ вказує на розрахунок довірчого інтервалу на основі бутстрепу.

- **Intervals**: Розділ показує довірчий інтервал для оцінки пропорції на рівні довіри 95%. Інтервал (0.4167, 0.5833) означає, що на основі даних та виконаного аналізу, можемо бути впевнені з 95% вірогідністю, що реальна пропорція вижившиx жінок в третьому класі "Титаніку" знаходиться в межах від приблизно 41.67% до 58.33%.
- **Calculations and Intervals on Original Scale**: Цей розділ вказує, що розрахунки і довірчий інтервал проведені на вихідній шкалі (оригінальній шкалі), тобто для оцінки пропорції "так" у вибірці **women_third_class**.

ВИСНОВОК

Протягом нашого дослідження ми розглянули основи неймовірнісних вибірок і способи їх використання для оцінки пропорцій. Неймовірнісні вибірки є важливим інструментом в дослідницькій практиці, особливо коли вимоги до випадкової вибірки не можуть бути виконані

Ми розглянули різні форми неймовірнісного відбору, такі як метод типових представників, квотна вибірка, вибірка "гніздова", метод "снігової груди" та стихійна вибірка. Кожен з них має свої особливості та може бути ефективним в різних контекстах.

Застосування неймовірнісних вибірок для оцінки пропорцій - це ще одна важлива область, яку ми детально розглянули. Ми розглянули три основні методи: метод підстановки, метод зважених оцінок та метод бутстреп. Кожен з цих методів було продемонстровано на практиці за допомогою мови програмування R.

Ці методи, хоча і мають свої переваги та недоліки, можуть служити важливими інструментами для використання неймовірнісних вибірок. Їх застосування може допомогти отримати корисні інсайди з даних, коли звичайні методи вибірки неможливі або недоступні.

Загалом, дослідження показало, що неймовірнісні вибірки та асоційовані з ними методи оцінки пропорцій представляють значний потенціал для дослідження різноманітних важливих питань. Вони продовжують бути важливим інструментом для дослідників в різних галузях.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Rubin, Donald B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
2. Little, Roderick JA; Rubin, Donald B. (2002). Statistical Analysis with Missing Data (Second ed.). Hoboken, NJ: Wiley.
3. Van Buuren, S. (2018). Flexible Imputation of Missing Data. Second Edition. Chapman & Hall/CRC. Boca Raton, FL.
4. Groves, R. M. (2009). Survey Methodology. John Wiley & Sons.
5. Efron, B. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC.
6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.
7. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.
8. Wickham, H., & François, R. (2016). dplyr: A grammar of data manipulation. R package version 0.5.0.
9. Grolemund, G., & Wickham, H. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.
10. Canty, A., & Ripley, B. (2020). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25. Davison, A.C. & Hinkley, D.V. (1997) Bootstrap Methods and Their Application. Cambridge University Press.
11. Jean-François Beaumont (2020) Are probability surveys bound to disappear for the production of official statistics?
12. V. Nekrasaite-Liege & A. Ciginas1,3 & D. Krapavickaite (2022) Usage of non-probability sample and scraped data to estimate proportions.
13. Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society.

14. Rao, C. R., & Molina, I. (2015). Small Area Estimation (2nd Edition). John Wiley & Sons.

ДОДАТОК

2.1.1.1 Середнє значення

```
# Встановлюємо пакет, якщо він ще не встановлений
if (!require(mice)) {
  install.packages("mice")
}
# Завантажуємо пакет
library(mice)
# Створюємо простий набір даних з пропущеними значеннями
data <- data.frame(age = c(25, 27, NA, 30, 32, 35, NA, 40, 43, 47, NA),
        gender = c("M", "F", "M", "F", "M", "F", "M", "F", "M", "F", "M"))
# Застосовуємо метод підстановки
imputed_data <- mice(data, m = 1, method = "mean")
# Виводимо оригінальні та заповнені дані
print(data)
print(complete(imputed_data))
# Створюємо вікові групи
age_groups <- cut(complete(imputed_data)$age, breaks = c(0, 30, 40,
Inf), labels = c("Young", "Middle", "Old"))
# Розраховуємо пропорцію вікових груп
age_table <- table(age_groups)
age_proportion <- age_table / sum(age_table)
print(age_proportion)
```

2.1.1.2 Лінійна регресія

```
# Завантажуємо пакет
library(mice)
# Створюємо простий набір даних з пропущеними значеннями
data <- data.frame(
  age = c(25, 27, NA, 30, 32, 35, NA, 40, 43, 47, NA),
  income = c(30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000)
)
# Створюємо копію даних для роботи
data_imputed <- data
# Застосовуємо лінійну регресію для прогнозування пропущених значень "age" на основі "income"
model <- lm(age ~ income, data = data)
# Обчислюємо прогнозовані значення "age" для пропущених значень
predicted_age <- predict(model, newdata = data)
# Замінюємо пропущені значення "age" прогнозованими значеннями в копії даних
data_imputed$age[is.na(data_imputed$age)] <- predicted_age[is.na(data_imputed$age)]
# Виводимо оригінальні та заповнені дані
print(data)
print(data_imputed)
# Створюємо вікові групи
age_groups <- cut(data_imputed$age, breaks = c(0, 30, 40, Inf), labels = c("Young", "Middle", "Old"))
```

```
# Розраховуємо пропорцію вікових груп
age_table <- table(age_groups)
age_proportion <- age_table / sum(age_table)
print(age_proportion)
```

2.1.1.3 Множинна підстановка

```
# Завантажуємо пакет
library(mice)
# Створюємо простий набір даних з пропущеними значеннями
data <- data.frame(
  success = c(1, 0, 1, NA, NA, 0, 1, NA, 0, 1),
  group = c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)
)
# Застосовуємо множинну імпутацію
imp_data <- mice(data, m = 5, method = "pmm")
# Створюємо функцію для розрахунку пропорцій успіху
prop_fun <- function(data) {
  with(data, tapply(success, group, function(x) sum(x) / length(x)))
}
# Розраховуємо пропорції успіху для кожного набору даних, що імпутується
props <- apply(imp_data$imp$success, 2, function(x) {
  temp_data <- data
  temp_data$success[is.na(temp_data$success)] <- x
  prop_fun(temp_data)
})
```

```
# Розраховуємо середню пропорцію і стандартну помилку
props_mean <- mean(props)
props_se <- sd(props) / sqrt(length(props))
# Виводимо результати
print(data.frame(Proportion = props_mean, SE = props_se))
```

2.1.2 Метод зважених оцінок

```
# Завантажуємо потрібну бібліотеку
library(dplyr)
# Завантажуємо дані з CSV-файлу
data <- read.csv('/Users/marko/Desktop/R/survey.csv')
# Обчислюємо вагу для кожного спостереження виходячи з віку
data$weight <- data$Age / sum(data$Age)
# Перевіряємо, що ваги правильно розраховані
print(sum(data$weight))  # має бути приблизно 1
# Розраховуємо зважені пропорції відповідей "Yes" для кожної групи за статтю
data %>%
  group_by(Gender) %>%
  summarise(prop_yes = sum(weight[Response == "Yes"]) / sum(weight))
# Обчислюємо нову вагу , яка тепер враховує як вік , так і дохід
data$weight <- (data$Age + data$Income) / (sum(data$Sge) + sum(data$Income))
# Перевіряємо , що ваги правильно розраховані
print(sum(data$weight))  # має бути приблизно 1
```

```
# Обчислюємо зважену пропорцію відповідей "Yes" для кожної групи за статтю
data %>%
  group_by(Gender) %>%
  summarise(prop_yes = sum(weight[Response == "Yes"]) / sum(weight))
```

2.1.3 Метод Бутстреп

Приклад 1

```
# Підключаємо потрібні бібліотеки
library(tidyverse)
library(boot)
# Завантажуємо набір даних Titanic з Kaggle
titanic <- read.csv('/Users/marko/Desktop/R/titanic.csv')
# Створюємо підмножину даних з жінками, що подорожували третім класом
women_third_class <- titanic %>%
  filter(Sex == 'female' & Pclass == 3)
# Визначаємо функцію для вибірки з відновленням і розрахунку пропорції вижившихboot_func <- function(data, indices) {
  sample_data <- data[indices, ]
  return(sum(sample_data$Survived) / nrow(sample_data))
}
# Використовуємо метод бутстрепу з 5000 повтореннями
results <- boot(data = women_third_class, statistic = boot_func, R = 10000)
```

```
# Виводимо результати
print(results)
# Обчислюємо 95% довірчий інтервал
boot.ci(results, type="perc")
```