ЧЕРНЯК ОЛЕКСАНДР ІВАНОВИЧ

МЕТОДИ ПРОВЕДЕННЯ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ВИБІРКОВИХ ДОСЛІДЖЕНЬ

ЗМІСТ

РОЗДІЛ 1

ВИБІРКОВІ ДОСЛІДЖЕННЯ ТА ПОВ'ЯЗАНІ З НИМИ ПРОБЛЕМИ

1.1. *Основні ідеї вибіркових досліджень та оцінювання.*

1.2. *Елементи вибірки.*

1.3. *Вибіркові та невибіркові похибки.*

1.4. *Перевага вибіркових методів та можливі сфери використання техніки вибіркових досліджень.*

1.5. *Основні проблеми вибіркового дослідження.*

Майже всі статистичні методи базуються на простих випадкових вибірках. Термін вибіркове обстеження використовується в зв’язку з вибірками популяцій: груп людей, домогосподарств, фірм і таке інше. Сутність вибіркового методу дослідження (обстеження) полягає у відборі з сукупності досліджуємого матеріалу такої її частини (вибірки), яка повинна представити всю сукупність.

Надалі популяцією або генеральною сукупністю будемо називати сукупність одиниць або елементів, про які бажаємо зробити певні висновки. Із всієї популяції здобувається вибірка одиниць (вибіркові одиниці), по яким робляться висновки про властивості всієї популяції. Наприклад, щоб з’ясувати точку зору населення по тому чи іншому питанню, проводиться опитування лише частини населення, і отримані відповіді служать для оцінки пропорцій різних думок у всьому населенні. Щоб оцінити ступінь розповсюдження якогось захворювання, у вибірку беруть деяку кількість медичних закладів, і дивляться, скільки пацієнтів мають дану хворобу. Рейтинг кандидата на виборах у даному регіоні оцінюється на основі опитування порівняно невеликого числа виборців із різних районів даного регіону, і тому подібне.

Прикладом вибіркових обстежень може бути визначення середнього рівня доходів населення, визначення переліку споживчих переваг, оцінювання економічного потенціалу регіону, аналіз діяльності малих підприємств, оцінювання вартості робочої сили, оцінювання фінансово-економічних показників підприємств з

метою запобігання банкрутства, визначення кількості жителів регіону або країни, визначення рейтингу кандидата на виборах і таке інше.

В процесі вибіркового обстеження виникають такі запитання: як найкращим чином здійснити вибірку і (коли вибіркові дані вже отримані) як найкраще їх використати для адекватної оцінки характеристик популяції. При цьому потрібно вирішити, якими повинні бути обсяг вибірки, спосіб її отримання, методи спостережень. Отримання оцінок з мінімальними похибками пов'язане з вибором типу оцінки, можливістю застосування додаткової інформації з вибірки.

Вибіркові дослідження відрізняються від досить близької сфери експериментальних досліджень тим, що в експериментах дослідник впливає певним чином на частину сукупності, щоб побачити, до яких наслідків це призведе. У вибіркових дослідженнях передбачається тільки виявлення певних характеристик самої сукупності, без жодних впливів на неї. Тобто, дослідник сподівається, що формулювання питань ніяк не вплине на відповіді респондентів, або обстеження певних тварин у регіоні не призведе до зміни їх перебування чи поведінки.

Вибіркові дослідження також відрізняються від чисто дослідницької діяльності, в якій дослідника не цікавить спосіб отримання даних. У вибіркових дослідженнях кожен має можливість обрати специфічний метод, який по можливості дозволить уникнути багатьох факторів, що роблять вибіркові дані навмисними, зручними, "нерепрезентативними".

1.1. Основні ідеї вибіркових досліджень та оцінювання

Нехай популяція, яку треба дослідити, складається із скінченної кількості N одиниць – наприклад, людей чи земельних ділянок, яким можна присвоїти номери *1, 2, ... , N*. При цьому Y_i - значення або характеристика i - ї одиниці (**ознака**).

З усієї генеральної сукупності обирається та досліджується тільки вибіркова частина її одиниць у кількості *n*. Отримані дані складаються з *Y*-значень елементів, що потрапили до вибірки, разом з їх номерами; або, у загальному випадку, ще з декількома додатковими значеннями. Наприклад, при дослідженні точки зору населення, як додатки значення можуть бути: стать, вік чи величина доходу респондентів. Процедура, за допомогою якої окремі одиниці обираються з генеральної сукупності, носить назву тип відбору. Більша частина відомих типів відбору визначаються шляхом присвоєння кожній можливій вибірці *s* імовірності її обрання *P(s)*. Наприклад, для простого випадкового відбору з обсягом вибірки *n*, можлива вибірка складається з *n* окремих одиниць сукупності, а імовірність *P(s)* – однакова для кожної можливої вибірки *s*. На практиці тип відбору можна описати як покрокову процедуру вибору одиниць, на відміну від розгляду ймовірностей обрання окремої вибірки. При простому випадковому відборі така покрокова процедура полягає у випадковому виборі номера одиниці з {*1, 2, ... , N*}, випадковому виборі наступного номера елемента із номерів, що залишилися, і так далі, доки не буде обрано *n* різних елементів.

Вся послідовність *Y*-значень генеральної сукупності $Y_1, Y_2, ..., Y_N$ вважається її фіксованою характеристикою або параметром. В більшості випадків основною проблемою вибіркового дослідження є оцінка деякої сумарної характеристики популяції, такої як середнє або сумарне значення величин *Y*, маючи тільки вибіркові результати. Крім цього, у більшості випадків дослідника цікавить точність, надійність отриманих оцінок.

Якщо розмір вибірки розширити до включення абсолютно всіх N одиниць сукупності, то характеристика сукупності визначатиметься в точності. Таким чином,

ненадійність оцінок, отриманих з вибірки, основана на тому, що була досліджена тільки частина сукупності. Реальна характеристика сукупності залишається незмінною в той час як її оцінка залежить від обраної вибірки. Якщо для кожної можливої вибірки оцінка виявляється досить близькою до реального значення, тоді лише незначна неточність пов'язана із стратегією вибору; така стратегія – бажана. Якщо, з іншого боку, оцінка дуже різниться серед можливих вибірок, то неточність пов'язана із методом вибору. Особливістю найбільш корисних типів відбору є те, що ця різниця від однієї вибірки до іншої оцінюється за допомогою лише однієї окремої вибірки.

Приділяючи особливу увагу типові відбору та застосовуючи відповідний метод оцінювання, можна побудувати незміщені оцінки для таких характеристик сукупності, як середнє та сумарне значення, не вдаваючись до припущень щодо самої сукупності. Оцінка – незміщена, якщо її очікуване значення по всім можливим вибіркам згідно даного типу відбору (математичне сподівання) дорівнює істинному значенню сукупності. Крім цього, випадкове або ймовірне обрання вибірок дозволяє уникати відхилень, пов'язаних з людськими схильностями, такими як свідома чи несвідома тенденція до обрання елементів з більшими (або меншими) за середнє значеннями цікавлячої нас характеристики. Ця процедура особливо бажана, коли результатами досліджень користуватимуться особи з протилежними інтересами – наприклад, передвиборне визначення рейтингу кандидатів на посаду мера різними виборчими штабами. В таких випадках малоймовірно, щоб усі зацікавлені особи могли б домовитись про навмисний вибір "репрезентативної" вибірки.

Таким чином, імовірнісний відбір, такий як простий випадковий, може забезпечити отримання незміщених оцінок середнього чи сумарного значень сукупності, а також незміщену оцінку дисперсії, яка використовується для оцінювання правдоподібності або точності отриманого результату. Незміщені оцінки та оцінки дисперсії можна також отримати при використанні відборів, де одиниці потрапляють до вибірки не рівноможливо, а з певними ймовірностями, заданими наперед для кожної одиниці та для пар одиниць.

Поряд із метою одержання незміщених чи майже незміщених оцінок при дослідженні постають задачі отримання точних оцінок або оцінок з невеликою дисперсією та використання нескладних чи економічно ефективних процедур відбору. Бажання задовольнити якомога більше з цих завдань, знаходячись під впливом багатьох обставин, призвело до розвитку широко відомих типів відбору та методів оцінювання, включаючи простий випадковий відбір і відбір з нерівними імовірностями, застосування додаткової інформації, стратифікований, систематичний, багатоетапний і подвійний відбори, а також інші технічні прийоми.

1.2. Елементи вибірки

На початку відбору дуже важливо визначити тип одиниць, які слід обирати, їх множину згідно поставленої проблеми. Одиницями можуть бути люди, домашні господарства, ферми, лікарні чи підприємства. Їх повний список у сукупності, що досліджується, забезпечить ідеальну базу, з якої можна робити вибірку. На практиці інколи буває неможливим скласти повний перелік одиниць генеральної сукупності. Наприклад, множина усіх телефонних номерів, які можна набирати випадковим чином, не враховує тих сімей, які не мають телефонів; загальні або окремі списки можуть бути застарілими і т.п.

У деяких сукупностях буває досить важко визначити характер одиниць. При вивченні природних ресурсів або врожаю регіон може бути поділений на декілька географічних одиниць (сегментів), які потраплятимуть до вибірки за допомогою карти. Однак розмір та форма цих сегментів можуть бути довільними, і рішення дослідника щодо них може вплинути на вартість дослідження та точність оцінок.

Таким чином, побудова повного переліку одиниць відбору є однією із найважливіших проблем техніки вибіркових досліджень і є основою вибірки.

Слід також зауважити, що сукупність з якої робиться відбір (основа вибірки або обстежувана сукупність) повинна співпадати з сукупністю, про яку хочемо зібрати відомості (досліджувана сукупність).

Вкажемо на основні вимоги до вибіркової та генеральної сукупності. При формуванні вибіркової сукупності необхідно дотримуватися наступних умов [2, 5, 6, 9, 13] :

1. Кожна одиниця генеральної сукупності повинна мати однакову можливість попасти у вибірку.

2. Вибіркова сукупність повинна бути типовою, представницькою (репрезентативною) по відношенню до генеральної сукупності, відтворювати особливості генеральної сукупності.

3. Вибіркова сукупність повинна формуватися по принципами теорії ймовірностей.

4. Відносна однорідність вибіркової сукупності або поділ вибірки на однорідні групи одиниць спостереження.

5. Усі дані, які збираються по вибірковій сукупності, повинні відповідати меті проведеного обстеження;

6. Необхідно чітке визначення одиниці відбору;

7. Основа вибірки повинна бути точною, повною і вільною від подвійного рахунку і відповідати меті дослідження.

8. Необхідно враховувати фактори ротації вибірки. Так, основа вибірки, яка була точною, повною і вільною від подвійного рахунку в момент її побудови, може застаріти до того часу, коли нею потрібно буде користуватися. Наприклад, при обстеженні ділової активності промислових підприємств важливо врахувати, що протягом декількох років можуть змінитися як розмір підприємства, так і їх число: частина збиткових підприємств буде закрита, також можуть бути створені нові підприємства.

9. Визначення цензу для включення дрібних одиниць у вибіркову сукупність.

Вимоги, які ставляться до генеральної сукупності, менш жорсткіші:

1. Генеральна сукупність, досліджувана сукупність і об’єкт статистичного спостереження повинні бути ідентичними.

2. Генеральна сукупність повинна мати достатньо великий обсяг.

3. Необхідне чітке формулювання правил відбору із генеральної сукупності одиниць відбору.

4. Кожен елемент генеральної сукупності повинен належати тільки одній одиниці відбору.

5. Однорідність генеральної сукупності або поділ її на однорідні групи (страти).

6. У випадку реєстрації об’єктів тієї чи іншої генеральної сукупності необхідна їх регулярна актуалізація: облік змін розміру і основного виду діяльності об’єктів, а також структурних змін (утворення нових і ліквідація існуючих об’єктів).

1.3. Вибіркові та невибіркові похибки

Чи будуть результати вибірки достатньо репрезентативними для всієї сукупності , залежить від похибок, що вносяться самим процесом відбору. З точки зору традиційних вибіркових досліджень характеристика об’єктів вибірки вимірюється абсолютно точно, тому похибки в оцінках виникають тільки завдяки тому, що вибірка містить не всі одиниці генеральної сукупності. Ці похибки отримали назву “вибіркові похибки”. Їх середня величина залежить від обсягу вибірки, варіації ознак одиниць досліджуваної сукупності, від прийнятої методики відбору і способу обчислення шуканих величин.

Імовірнісна випадкова похибка є величиною, контрольованою за допомогою процедури складання випадкової вибірки. Цю похибку можна зменшити чи піддати кращому контролю двома способами: шляхом підвищення обсягу вибірки або поліпшення структури вибірки шляхом стратифікації чи кластеризації, які враховують варіацію в генеральній сукупності.

При правильному відборі середню величину випадкових похибок вибірки і навіть очікувану частоту появи похибок тої чи іншої величини можна обчислити по даним фактично отриманої вибірки. Усі способи обчислень базуються на математичній теорії вибіркового методу.

Результати аналізу, пов’язаного з обчисленням похибок вибірки, використовуються для оцінки ефективності різних способів відбору і для кращого планування подальших обстежень.

Але в реальних ситуаціях невибіркові похибки також мають місце. Деякі люди, наприклад, можуть відмовитись відповісти на питання, і такі відмовлення можуть бути нетиповими для населення, що вивчається. Тоді дані вибірки виявляться нетиповими для даного населення, а оцінки – зміщеними. Проблема відмовлення від відповіді особливо важлива при дослідженнях, коли люди взагалі відповідають мало, в яких імовірність відповіді пов'язана з самою характеристикою, що вивчається. Яскравим прикладом цього є вивчення сексуальної сторони життя людей. Ефект проблеми "не відповіді" (не отримання даних) можна пом'якшити завдяки додатковим намаганням оцінити характеристику тих респондентів, які не відповіли, використовуючи додаткову інформацію всіх респондентів або вдаючись до моделювання цієї ситуації. Але все-таки слід прикладати всі сили, щоб "не відповідей" було якомога менше.

Для визначення способів розв'язання проблеми "не відповіді" важливо виділити основні види не отримання інформації по деяким одиницям сукупності. В роботах [5, 13] пропонується приблизна класифікація видів "не відповіді":

1. Не виявлені. До них відносяться одиниці вибірки, які неможливо знайти або відвідати.
2. Відсутні вдома.
3. Нездатні дати відповідь. Респондент може не мати відомостей по деяким питанням або може не хотіти їх повідомити.
4. "Кріпкі горішки". Включають осіб, які категорично відмовляються відповідати, не в змозі відповісти або знаходяться далеко від дому протягом всього періоду обстеження. Вони є джерелом зміщення оцінок, яке не можна ліквідувати.

Причини "не відповіді" залежать від багатьох факторів, таких, як тип обстеження (по пошті, за допомогою інтерв'ю, шляхом безпосередніх вимірів і таке інше), тип респондента (усі люди в будинку чи одна людина, всі сім'ї чи ті, що належать до певної групи за доходами), кваліфікація і старанність інтерв'юера, привабливість опитуваної анкети, зміст питань, час проведення обстеження, кількість і зміст повторних спроб обстеження.

При розв’язанні проблеми “не відповіді” виділяють ряд методів. Їх можна об’єднати у дві великі групи:

1) способи зменшення і запобігання втрати даних;
2) методи мінімізації небажаних ефектів втрати даних.

Перший напрямок (повторне відвідування початкових одиниць, додаткові вибірки, повторне звернення) є найбільш ефективною політикою скорочення похибок спостережень. Вона зв’язана, в першу чергу, із збільшенням часових, грошових і трудових витрат до тих пір, поки необхідні дані не будуть отримані. При обході території число неохоплених елементів може бути зменшено при використанні гарних інструкцій і допоміжних засобів (карти, фотографії); для поштових і телефонних обстежень – при складанні точних списків адрес і номерів телефонів, для усіх випадків - при залученні підготовлених фахівців і розробці зручних схем виявленні помилок. Число осіб, які “відсутні вдома”, можна зменшити за допомогою повторних візитів, а число “відмов” – за допомогою привабливих питань, майстерно проведеного інтерв’ю, розумно складених опитувальних анкет. Повернення чи повторне відвідування можуть бути економічними, якщо вірно визначити ймовірність того, що респондент виявиться вдома.

Методи мінімізації небажаних ефектів втрати даних включають в себе:

- спосіб визначення оптимальної частки відбору серед тих респондентів, які не відповіли при вибірковому обстеженні ;
- поправку на зміщення без повторних звернень;
- обстеження з двома зверненнями;
- оцінка втрачених значень;
- методи оцінки характеристик тих респондентів, які не відповіли при вибірковому обстеженні.

Серед перерахованих способів усунення похибок спостережень найбільш розробленими і надійними є: повторне звернення, визначення оптимальної частки відбору серед тих респондентів, які не відповіли і методи оцінки їх характеристик .

При оцінці характеристик тих респондентів, які не відповіли при вибірковому обстеженні, або при відсутності даних вибіркового обстеження застосовується процедура імпутації.

Імпутація є одним з важливих етапів обробки даних вибіркового обстеження. Імпутація - це процедура заповнення відсутніх значень по окремих ознаках або по групах ознак, які вимірюються за програмою обстеження [7]. Ця процедура застосовувалась НДІ статистики Державного комітету статистики України для обстеження умов життя домогосподарств [4].

За наявністю будь-яких відсутніх значень (пропусків) в даних вибіркового обстеження є два шляхи їх компенсації при обробці результатів, а саме - відкидання спостережень з відсутніми значеннями і перезважування даних по одиницях вибірки, що залишилися, або штучне заповнення пропусків - імпутація відсутніх даних. На практиці зважування здійснюють для врахування випадків, коли по окремих одиницях вибірки взагалі немає інформації або немає суттєвих даних, а імпутацію - коли немає лише окремих даних. Таким чином, однією з головних умов для можливості здійснення процедури імпутації вважається наявність деякої суттєвої інформації по одиницях вибірки.

Найбільш поширеною групою методів аналізу даних з пропусками є так звані методи з заповненням. При використанні цих методів пропуски заповнюються величинами, які визначаються за допомогою спеціальних процедур, і отримані “повні” дані оброблюються стандартними статистичними методами. Для заповнення використовуються такі основні процедури [7]:

- заповнення середніми або метод “середнього значення” (замість пропущених величин підставляються середні, розраховані по присутніх даних вибіркового обстеження);

- заповнення за пропорцією або метод “пропорцій” (замість пропущених величин підставляються величини, визначені з умови збереження пропорції присутніх даних);

- заповнення з підбором, зокрема метод “hot-deck” (підставляються значення змінних інших об’єктів вибірки).

Методи заповнення мають ту перевагу, що вони одночасно дають “повну” матрицю даних і дозволяють використати практично всі отримані в обстеженні дані, оскільки не передбачають відкидання спостережень з пропусками у значеннях окремих ознак.

При виборі методу імпутації слід мати на увазі, що якщо пропущених даних небагато, наприклад, близько 1 чи 2 відсотки, то практично немає значення, який метод застосовується для заповнення пропусків. Використання будь якого методу призведе до тих ж самих результатів. Але в більшості вибіркових обстежень домогосподарств, рівень пропусків є значно вищим, і досягає інколи 30 відсотків для окремих ознак [4]. При такому рівні пропусків ефектом впливу штучних значень показників вже не можна нехтувати. Невдале заповнення пропусків може призвести до виникнення систематичних похибок (зміщення) в оцінках ознак. У зв’язку з цим перед застосуванням процедури імпутації слід проводити перевірку коректності обраного методу імпутації відсутніх даних. Крім того, необхідно завжди позначати ті значення ознак, які отримані штучно. Це дозволить користувачам, в разі необхідності, більш адекватно уявити собі ступінь надійності даних, застосувати іншу процедуру імпутації тощо. При комп’ютерній обробці даних імпутовані значення ознак зручно позначати за допомогою спеціальних ознак - так званих змінних-“прапорців”. Змінна-“прапорець” - ознака, яка набуває одного значення для наявних даних (наприклад, “1”) і іншого значення для даних, які було імпутовано (наприклад, “2”).

Також причиною невибіркових похибок є похибка обліку (складання списку). Вона виникає через неточність в базі обстеження (опитування), тобто невідповідності між реальними об’єктами досліджуваної сукупності і їх списком. Об’єкти, які належать до генеральної сукупності, можуть виявитися не внесеними до списку, і навпаки, неіснуючі об’єкти можуть фігурувати в базі обстеження.

При вимірюванні, обробці та збереженні інформації можуть також з’являтися похибки. Якість процедури вимірювання (збирання) інформації залежить від методів роботи опитувача “в полі” і надійності отриманих відповідей. Якість процедури обробки інформації залежить насамперед від дослідницької фірми, яка займається

обробкою й аналізом даних. Тому для зведення таких похибок до мінімуму важливо забезпечити якісний контроль на кожному етапі дослідження. В деяких випадках вдається змоделювати похибки вимірювання окремо від результатів досліду, щоб пов'язати дані спостереження з реальними характеристиками сукупності [23].

Взагалі, проведення суцільного обстеження сукупності вимагає багато часу і грошей і не ліквідує похибку.

Суцільне обстеження може бути доцільним , коли досліджувана сукупність мала, що часто буває у дослідженнях промислових ринків, де кількість підприємств в обстежуваному секторі 100-300, і суцільне обстеження може бути реальним і з точки зору витрат, і з точки зору швидкості отримання інформації. Але навіть у ситуаціях такого типу суцільне обстеження може бути справді корисним, тільки якщо досліджувані змінні варіюють у достатній мірі. Якщо всі досліджувані одиниці сукупності ідентичні з точки зору досліджуваної змінної, то суцільне обстеження буде не чим іншим, як марнуванням коштів і даремною витратою часу в ситуації, коли достатнім було б обстеження однієї чи декількох одиниць (респондентів). У цьому виявляється одне загальне і важливе правило визначення обсягу вибірки : чим більша варіація досліджуваного фактора, тим більшою має бути вибірка.

1.4. Переваги вибіркових методів та можливі сфери використання техніки вибіркових досліджень

Розробка перерахованих прийомів дослідження дозволила перетворити вибірковий метод у надійний спосіб статистичного спостереження.

Можна виділити **три головні переваги** використання вибіркового методу.

1. Застосування техніки вибіркових досліджень може забезпечити надійну інформацію по значно меншій собівартості, ніж суцільне обстеження всієї сукупності (наприклад, суцільний перепис населення).

2. Дані можуть бути зібрані значно швидше. Оцінка рівня безробіття у 2003 році не дуже корисна, якщо опитування всіх сімей і обробка результатів буде зроблена до 2005 року.

3. Нарешті, оцінки, отримані з використанням вибіркового методу, є часто більш точні, ніж ті, що базуються на суцільному обстеженні популяції, через те, що вибіркові обстеження можуть проводити більш кваліфіковані кадри. Суцільний перепис населення часто вимагає велику адміністративну організацію і включає багато осіб (більшість - некваліфікованих) у збирання даних. З адміністративною складністю і тиском, щоб своєчасно зробити оцінки, з'являються похибки при суцільному перепису населення. При вибірковому обстеженні, більше уваги може бути приділено якості даних, проблемам "не відповіді" із залученням навченого кваліфікованого персоналу. Значно краще мати якісні вимірювання на репрезентативній вибірці, ніж ненадійні чи зміщені вимірювання на всій популяції.

У.Демінг (W.Deming) у своїй книзі "Some Theory of Sampling" пише, що " Техніка вибіркових досліджень - це наука і мистецтво управління і вимірювання надійності корисної статистичної інформації з використанням теорії ймовірностей" [16].

Можливі сфери використання техніки вибіркових досліджень:

1. Вибіркові обстеження з метою отримати інформацію, необхідну для прогнозу національної економіки. Теми обстеження: промислове виробництво, торгівля, сільськогосподарське виробництво та землекористування, безробіття та зайнятість, оптові та роздрібні ціни, стан здоров'я людей, доходи та витрати сімей.

2. Вибіркові обстеження по частковим темам: житлові та соціальні проблеми літніх людей; придбання товарів довгострокового використання; заборгованість орендаторів; вартість житлового будівництва; вплив телебачення на школярів; зайнятість вчених і інженерів у промисловості; обстеження смаку та стандартів харчування і таке інше.

3. Використання вибіркових обстежень при перепису населення: отримання додаткової інформації про зайнятість, походження, дітях, доходах, освіті і таке інше.

4. Використання вибіркових досліджень на рівні місцевої влади: прогнозування розвитку економіки міста; розв'язання соціально-економічних проблем; житлове обстеження; показники забрудненості навколишнього середовища; рівень захворюваності населення міста і таке інше .

5. Вибіркове дослідження ринку: фінансове обстеження споживачів; оцінка радіо- і телепрограм; обстеження читацької аудиторії; аналіз споживчого попиту; ринок житла та житлових умов і таке інше .

6. Галузеві вибіркові обстеження з метою підвищення ефективності резидентів (підприємств, організацій): основні показники комерційної діяльності підприємств, організацій; тенденції в розвитку виробництва, попиту, пропозиції продукції, динаміки цін; запаси матеріалів і готової продукції; інвестиції; використання виробничих потужностей.

7. Вибіркові обстеження окремих партій виробів, окремих робочих періодів: якість і прийом продукції; використання робочого часу; використання фонду робочого часу обладнання і таке інше.

8. Опитування громадської думки і передвиборні опитування: думки про соціально-економічну ситуацію в країні, відношення до уряду; відношення до політичних проблем, партій; до судових процесів і таке інше.

1.5. Основні проблеми вибіркового дослідження

В практиці використання техніки вибіркових досліджень можна виділити два типа проблем: загальні та специфічні.

Загальні обумовлені самим характером вибірки як окремого агрегованого методу статистичних досліджень. В залежності від етапів проведення вибіркових обстежень доцільно визначити ряд вузлових проблемних питань, від розв’язку яких залежить ефективність реалізації вибірки:

- чітке формулювання мети обстеження;
- виконання всіх вимог, що пред’являються до генеральної сукупності, вибіркової сукупності, одиниці відбору;
- визначення бажаної точності; похибка вибірки зменшується із збільшенням обсягу останньої, але при цьому зростають часові та грошові витрати, що породжує необхідність вибору оптимальної ступені точності;

- вибір оптимального способу спостережень (інтерв'ювання, використання пошти, телеграфу, електронної пошти, Інтернету і таке інше), розробка форм опитуваних анкет, заключних таблиць;

- визначення оптимального способу відбору; знаючи бажаний рівень точності, для кожного способу відбору наближено оцінити обсяг вибірки;

- розробка методичного та програмного забезпечення;

- проведення пробного обстеження невеликого обсягу для перевірки методики обстеження;

- організація основного обстеження (навчання персоналу, контроль за їх роботою, перевірка якості відповідей, у випадку відсутності відповідей отримання інформації від окремих одиниць у вибірці);

- зведення та аналіз даних;

- використання інформації для наступних досліджень (кожне завершене вибіркове обстеження є потенційним засобом покращення наступного відбору, оскільки містить дані про середні значення, дисперсії основних характеристик, що дозволяє запобігти ряд помилок у наступних обстеженнях).

Специфічні проблеми пов'язані з особливостями розвитку статистики в Україні. Якщо в умовах командної економіки домінувала система суцільної звітності, то при формуванні ринкових відносин різко зросла потреба у використанні методів несуцільного статистичного спостереження , особливо вибіркових. Це обумовлено як виникненням великого числа малих підприємств і неможливістю їх суцільного обліку, так і деякою надлишковою інформацією, отриманою в результаті суцільного спостереження.

До головних проблем використання вибіркових методів відносять наступні [2, 8, 13] :

- перехід до широкого використання техніки вибіркових досліджень є комплексною проблемою і порушує чотири аспекти :

1) зміна організації статистики у частині збору первинної інформації;

2) підготовка фахівців;

3) технічне забезпечення;

4) розвиток наукових досліджень;

- досвід використання вибіркових методів в промисловості, торгівлі показав, що основні труднощі породжуються малими обсягами обстежуваних сукупностей при вимозі високої точності і при умові великого числа груп (по видам продукції, по території, по формам власності і таке інше); це веде до поділу генеральної сукупності на малі групи;

- необхідність спостереження за великим число показників висуває на одне з перших місць проблеми формування багатовимірних вибірок великого обсягу, для яких відсутні загальноприйняті методи;

- відсутність методичного забезпечення, сучасних монографій, перекладів зарубіжної літератури;

- відсутність досвіду використання вибіркових методів у персоналу, що проводить обстеження;

- можливості розповсюдження вибіркових обстежень скорочує та обставина, що діюча система статистичних показників перевантажена натуральними показниками і занадто деталізована за видами продукції.

Контрольні запитання, вправи

Контрольні запитання й завдання

1. *У чому полягає сутність імовірнісного відбору?*
2. *Які основні вимоги до вибіркової сукупності?*
3. *Які вимоги ставляться до генеральної сукупності?*
4. *Які шляхи розв'язання проблеми „не відповіді" при вибіркових обстеженнях?*
5. *Дайте визначення процедури імпутації.*
6. *Які головні переваги використання вибіркового методу?*
7. *Вкажить на можливі сфери використання техніки вибіркових досліджень.*
8. *Які виникають проблеми при використанні техніки вибіркових досліджень?*

9. *Охарактеризуйте специфічні проблеми вибіркових обстежень пов'язані з особливостями розвитку статистики в Україні.*

РОЗДІЛ 2

МЕТОДОЛОГІЧНІ ОСНОВИ ПРОСТОГО ВИПАДКОВОГО ВІДБОРУ ТА СИСТЕМАТИЧНОГО ВІДБОРУ

2.1. *Здобуття простої випадкової вибірки.*

2.2. *Оцінювання середнього та сумарного значень сукупност.і*

2.3. *Надійні інтервали.*

2.4. *Оцінювання обсягу вибірки.*

2.5.*Оцінювання пропорції сукупності.*

2.6. *Обсяг вибірки для оцінювання пропорції.*

2.7. *Оцінювання відношення.*

2.8. *Основні поняття систематичного відбору.*

2.9. *Оцінювання середнього та сумарного значення сукупност.*

Простий випадковий відбір є основною формою ймовірнісного відбору і забезпечує теоретичну основу для більш складних форм відбору. Можливі два шляхи здобуття простої випадкової вибірки: із поверненням, у якій одна й та ж одиниця може включатися більш, ніж один раз у вибірку; і без повернення, у якій всі одиниці у вибірці різні.

Проста випадкова вибірка з поверненням (SRSWR – simple random sampling with replacement) обсягу n із сукупності з N одиниць може розумітися, як n незалежних вибірок обсягу 1 із всієї сукупності. Кожна одиниця вибирається з ймовірністю $1/N$. Таким чином вибирається n одиниць, які можуть включати і дублікати від популяції. Однак, у скінченій генеральній сукупності відбір однієї і тієї ж одиниці двічі не забезпечує додаткової інформації. Тому звичайно віддається перевага вибірці без повернення. Проста випадкова вибірка без повернення (SRS – simple random sampling without replacement) обсягу n вибирається так, що всі комбінації з n різних одиниць в популяції мали однакові ймовірності бути

обраними. Оскільки, всього можливо C_N^n вибірок, то ймовірність вибору будь-якої індивідуальної вибірки *S* із *n* одиниць буде

$$P(S) = 1/C_N^n = \frac{n!(N-n)!}{N!}, \text{ де } n! = 1 \cdot 2 \cdots (n-1) \cdot n .$$

Таку вибірку можна отримати за *n* кроків, обираючи по одній одиниці за крок, при цьому кожного разу всі одиниці, що не обиралися, можуть рівноможливо потрапити до вибірки. При простому випадковому відборі без повернення (надалі – простий випадковий відбір) ймовірність того, що *i*-та одиниця генеральної сукупності потрапить до вибірки, дорівнює $\pi_i = n/N$, тобто вона однакова для всіх одиниць.

Інші типи відбору також можуть передбачити однакову ймовірність обрання для кожної одиниці, але тільки при простому випадковому відборі кожна можлива вибірка з *n* одиниць має однакову ймовірність обрання.

Дві різні прості випадкові вибірки обсягу *n*=9 із популяції *N*=100 показані на рис.1 та рис.2.

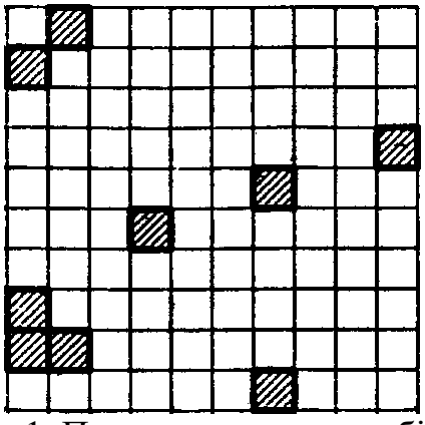

Рис. 1. Проста випадкова вибірка обсягу 9 із сукупності у 100 одиниць.

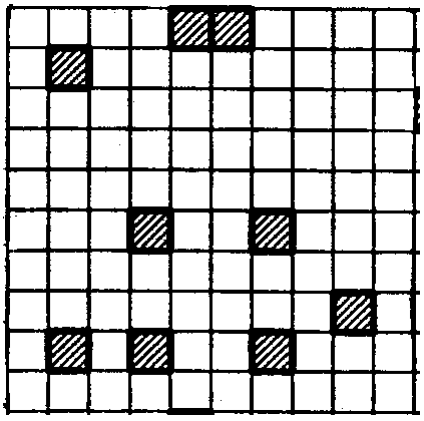

Рис. 2. Інша проста випадкова вибірка обсягу 9 із сукупності у 100 одиниць

2.1. Здобуття простої випадкової вибірки

Проста випадкова вибірка може бути вибрана, наприклад, за допомогою кульок з номерами від *1* до *N*, що містяться у ящику, і без погляду, вибираючи *n*

кульок без заміни будь-якої. Це робиться при проведенні лотерей чи конкурсів (спосіб жеребкування).

Щоб зменшити роботу процесу відбору, вибір робиться більш просто, використовуючи таблицю випадкових чисел або комп'ютерний датчик "випадкових чисел".

Наприклад, потрібно здобути просту випадкову вибірку обсягу n=10 із генеральної сукупності N=70 одиниць.

Скористаємось статистичною таблицею (див.додаток, табл.А.1). Вибираємо будь-яку колонку таблиці (наприклад, другу) і будь-яких два стовпчика цифр (наприклад, перші два). Рухаючись вниз по колонці, виписуємо 10 чисел, менших за 70: 6, 63, 62, 37, 29, 60, 27, 64, 24, 25. Це і є номери елементів, що потрапили до вибірки. Числа 72, 77, 75, 83, 90 були пропущені, так як вони більші за N.

Символ "00", коли використовуються два стовпчика, буде інтерпретовано, як одиниця 100. Якщо ж зустрічаються повтори, то число вибирається лише один раз. Наприклад, n=16, N=75, маємо 6, 69, 72, 62, 37, 29, 60, 75, 27, 64, 24, 25, 16, 19, 66, 59 (числа 72 та 27 були пропущені, оскільки вони вже присутні у вибірці).

Якщо N тризначне число (наприклад, N=152), тоді для прискорення процедури вибору можна скористатися таким методом: беремо тепер в таблиці будь-яку колонку і будь-яких три стовпчика цифр, якщо зустрічається тризначне число від 201 до 400, то віднімаємо від нього 200, якщо від 401 до 600 – віднімаємо 400, від 601 до 800 – віднімаємо 600, від 801 до 999 – віднімаємо 800. Всі числа, які більші 152 і менші 200, а також число 000, пропускаємо

Нехай n=8. Скористаємось четвертою колонкою. Маємо 64, 7, 26, 119, 67. 60, 10, 78.

Якщо $N \in [200;300)$, наприклад, N=282, то при появі чисел від 301 до 600 – віднімаємо 300, від 601 до 900 – віднімаємо 600. Числа, більші 900, пропускаємо.

Аналогічно для чисел $N \in [300;400)$, при появі числа від 401 до 800, віднімаємо 400, а числа, більші за 800, пропускаємо.

Якщо $N \in [400;500)$, то при появі чисел від 501 до 999, віднімаємо від них 500.

2.2. Оцінювання середнього та сумарного значень сукупності

Позначимо через $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$ - середнє значення сукупності. Будемо позначати через $y_1, y_2, ..., y_n$ значення одиниць, які потрапили у просту випадкову вибірку обсягу *n*.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ - вибіркове середнє.

Теорема 2.1 [9, 13]. Вибіркове середнє $\bar{y}$ є незміщеною оцінкою середнього значення сукупності μ.

Позначимо тепер через S^2 дисперсію всієї сукупності, тобто

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2 \tag{2.1}$$

Знайдемо дисперсію оцінки $\bar{y}$.

Теорема 2.2 [9, 13].

$$D\bar{y} = E(\bar{y} - \mu)^2 = \frac{S^2}{n}\frac{(N-n)}{N} = \frac{S^2}{n}(1-f), \tag{2.2}$$

де $f = \frac{n}{N}$ - частка відбору.

Величину $1-f = 1-\frac{n}{N}$ називають поправкою на скінченну сукупність.

Позначимо через $\sigma_{\bar{y}} = \sqrt{D\bar{y}}$ - стандартну похибку оцінки,

тоді

$$\sigma_{\bar{y}} = \frac{S}{\sqrt{n}}\sqrt{1-f} \tag{2.3}$$

Нехай тепер $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ - вибіркова дисперсія.

Теорема 2.3 [9, 13]. Статистика s^2 є незміщеною оцінкою дисперсії популяції S^2.

Наслідок 2.1. Незміщеною оцінкою дисперсії $D\bar{y}$ є

$$\upsilon(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{s^2}{n}(1-f) \tag{2.4}$$

Позначимо через τ - сумарне значення сукупності, тобто $\tau = \sum_{i=1}^{N} Y_i = N\mu$. З теореми 2.1 випливає, що незміщеною оцінкою τ буде статистика $\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i$.

Легко бачити, що

$$D\hat{\tau} = N^2 D\bar{y} = N(N-n)\frac{S^2}{n} = \frac{N^2 S^2}{n}(1-f) \tag{2.5}$$

і стандартна похибка

$$\sigma_{\hat{\tau}} = \frac{NS}{\sqrt{n}}\sqrt{1-f} \tag{2.6}$$

Незміщеною оцінкою $D\hat{\tau}$ буде статистика

$$\upsilon(\hat{\tau}) = \frac{N^2 s^2}{n}(1-f) \tag{2.7}$$

2.3. Надійні інтервали

Отримавши вибірку і використавши вибіркові дані для оцінки середнього чи сумарного значення сукупності, бажано зробити висновки відносно точності цих оцінок. Це найбільш часто робиться за допомогою побудови надійних інтервалів, в межах яких лежить істинне значення з досить великою ймовірністю. Нехай I – надійний інтервал для середнього значення сукупності μ. Вибираючи досить малою ймовірність похибки α, маємо $P\{\mu \in I\} = 1-\alpha$. Інтервал I називається $1-\alpha$-надійним інтервалом. Типові значення для похибки α є: 0,01; 0,05; 0,1. Нехай $\alpha = 0{,}05$, тоді $1-\alpha = 0{,}95$ і це означає, що з ймовірністю 0,95 істинне значення μ потрапляє в інтервал *I*.

Побудова $1-\alpha$-надійного інтервалу для μ та τ базується на нормальній апроксимації розподілу $\bar{y}$.

$\bar{y} - t_\alpha \sqrt{\upsilon(\bar{y})} < \mu < \bar{y} + t_\alpha \sqrt{\upsilon(\bar{y})}$, або

$$\bar{y} - t_\alpha \sqrt{(1-f)\frac{s^2}{n}} < \mu < \bar{y} + t_\alpha \sqrt{(1-f)\frac{s^2}{n}},$$

де t_α – табличне значення розподілу Стьюдента з $n-1$степеню свободи (див. таблицю А.2 додатка). Число t_α знаходиться в таблиці А.2 за ймовірністю $1-\alpha$ і числом степенів свободи $k = n-1$. При розмірах вибірки $n > 50$ можна використовувати стандартне нормальне наближення.

$$\bar{y} - c_\alpha \sqrt{\upsilon(\bar{y})} < \mu < \bar{y} + c_\alpha \sqrt{\upsilon(\bar{y})},$$

де c_α знаходиться з таблиці А.3 за ймовірністю $1-\alpha$.

Аналогічно будуються і надійні інтервали для сумарного значення τ.

$$\tau \in \left(\hat{\tau} - t_\alpha \sqrt{\upsilon(\hat{\tau})};\ \hat{\tau} + t_\alpha \sqrt{\upsilon(\hat{\tau})}\right), \text{ або } \tau \in \left(\hat{\tau} - t_\alpha N \sqrt{\frac{s^2}{n}(1-f)};\ \hat{\tau} + t_\alpha N \sqrt{\frac{s^2}{n}(1-f)}\right).$$

При $n > 50$ t_α можна замінити на c_α.

Приклад 2.1 Зібрано 676 підписних листів, на кожному з яких може бути до 42 підписів. Для оцінки загального числа підписів відібрали $n = 50$ листів ($\approx 7\%$ проста вибіркова вибірка). Дані про число підписів на цих 50 листах наведені у таблиці 2.1(x_i - число підписів, f_i - частота).

Таблиця 2.1

Дані про число підписів на листах

x_i	42	41	36	32	29	27	23	19	16	15	14	11	10	9	7	6	5	4	3
f_i	23	4	1	1	1	2	1	1	2	2	1	1	1	1	1	3	2	1	1

$$n = \sum f_i = 50,\ \ \bar{y} = \frac{1}{n}\sum f_i x_i = \frac{1471}{50} = 29{,}42,$$

$\hat{\tau} = N\bar{y} = 676 \cdot 29{,}42 \approx 19888$ підписів.

$\sum f_i x_i^2 = 54497$,

$$s^2 = \frac{1}{n-1}\left(\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{n}\right) = \frac{1}{49}\left(54497 - \frac{(1471)^2}{50}\right) = 229.$$

Нехай $1-\alpha = 0{,}8$, тоді $c_\alpha = 1{,}28$ і

$$\tau \in \left(19888 - \frac{1{,}28 \cdot 676}{\sqrt{50}} \cdot \sqrt{229(1-0{,}074)};\ 19888 + \frac{1{,}28 \cdot 676}{\sqrt{50}} \cdot \sqrt{229 \cdot (1-0{,}074)} \right).$$

Отже, $\tau \in (18107;\ 21669)$ з ймовірністю 0,8. Істинне значення τ було 21045.

2.4. Оцінювання обсягу вибірки

Одне з головних питань при плануванні вибіркового обстеження стосується обсягу вибірки для щоб досягнення бажаної точності.

Нехай максимально допустиме відхилення $\bar{y}$ від μ буде d і $P\{|\bar{y} - \mu| \geq d\} < \alpha$, де α - досить мала ймовірність.

Припустимо, що оцінка $\bar{y}$ має нормальний розподіл, тоді $P\left\{ \frac{|\bar{y} - \mu|}{\sigma_{\bar{y}}} \geq c_\alpha \right\} = P\{|\bar{y} - \mu| \geq c_\alpha \sigma_{\bar{y}}\} = \alpha$, або $P\{|\bar{y} - \mu| < c_\alpha \sigma_{\bar{y}}\} = 1 - \alpha$.

Таким чином, необхідно вибрати таке n, щоб $c_\alpha \sigma_{\bar{y}} = d$. Маємо $c_\alpha \sqrt{\left(\frac{N-n}{N} \right) \frac{S^2}{n}} = d$.

Розв'язуючи це рівняння відносно n, дістанемо

$$n = \frac{c_\alpha^2 S^2}{d^2 + \frac{c_\alpha^2 S^2}{N}} = \frac{n_0}{1 + \frac{n_0}{N}}, \qquad (2.8)$$

де $n_0 = \frac{c_\alpha^2 S^2}{d^2}$.

У випадку, коли оцінюється сумарне значення сукупності, маємо наступне рівняння $c_\alpha \sigma_{\hat{\tau}} = d$ або $c_\alpha \sqrt{N(N-n)\frac{S^2}{n}} = d$ і

$$n = \frac{n_0}{1 + \frac{n_0}{N}}, \text{ де } n_0 = \frac{N^2 c_\alpha^2 S^2}{d^2}.$$

Приклад 2.2. Нехай $N = 500$, $S^2 = 85$. Який повинен бути обсяг вибірки, щоб відхилення середнього вибіркового значення від істинного не перевищувало 2, $\alpha = 0{,}05$?

$$n_0 = \frac{c_\alpha^2 S^2}{d^2} = \frac{1{,}96^2 \cdot 85}{4} = 81{,}634 \approx 82,$$

$$n = \frac{82}{1 + \frac{82}{500}} \approx 70.$$

Розраховуючи обсяг вибірки на практиці можна зіткнутися з проблемами, оскільки дуже часто немає даних про оцінку дисперсії генеральної сукупності. У цій ситуації можливі три шляхи їх вирішення:

- якщо є можливість, то звернутися до результатів схожого дослідження, здійсненого в минулому, і використати значення дисперсії, оціненого в цьому досліджені;
- можна провести попереднє дослідження невеликої кількості (біля 30) одиниць і на цій основі оцінити дисперсію;
- якщо відомі максимальне $Y_{\max}$ і мінімальне $Y_{\min}$ значення досліджуваної ознаки в генеральній сукупності і гіпотеза про нормальний характер розподілу ознаки прийнятна значення середнього квадратичного відхилення (стандартної похибки) вираховується таким чином: $S \approx \frac{Y_{\max} - Y_{\min}}{6}$.

Зміст наведеної формули пояснюється так. Якщо гіпотеза про нормальний розподіл прийнятна, то дистанція між максимально і мінімально можливими значеннями випадкової змінної не перевищує $6S$ (правило 3σ) з ймовірність 0,9973 [10].

2.5. Оцінювання пропорції сукупності

В деяких ситуаціях при проведенні вибіркових досліджень необхідно оцінити пропорцію одиниць (або частку) в сукупності, які мають деяку властивість (атрибут). Наприклад, необхідно оцінити частку виборців, які підтримують даного

кандидата, або оцінити частку недержавних підприємств, які зайняті у сфері виробництва і т.п.

У такій ситуації, змінна, що характеризує i-ту одиницю Y_i є індикатором: $Y_i = 1$, якщо i-та одиниця має дану властивість, $Y_i = 0$, якщо це не виконується. Таким чином, $\tau = \sum_{i=1}^{N} Y_i$ - це число одиниць в сукупності, які мають певну властивість, а середнє значення сукупності μ - це пропорція (частка) одиниць в сукупності з даною властивістю. Позначимо через p пропорцію одиниць з атрибутом у сукупності, тобто $p = \frac{1}{N}\sum_{i=1}^{N} Y_i = \mu$. Дисперсія сукупності у цьому випадку буде такою:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - p)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - Np^2\right) = \frac{\left(Np - Np^2\right)}{N-1} = \frac{N}{N-1}p(1-p)..$$

Тут було використано те, що $\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} Y_i^2$, оскільки Y_i може приймати лише два значення: 0 або 1.

Позначимо тепер через $\hat{p}$ вибіркову пропорцію, тобто $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{\left(\sum_{i=1}^{n} y_i^2 - n\hat{p}^2\right)}{n-1} = \frac{n}{n-1}\hat{p}(1-\hat{p}).$$

Використовуючи формулу (2.2), маємо

$$D\hat{p} = \frac{S^2}{n}\left(\frac{N-n}{N}\right) = \frac{Np(1-p)}{n(N-1)}\cdot\frac{(N-n)}{N} = \frac{(N-n)}{(N-1)}\cdot\frac{p(1-p)}{n}.$$

Стандартна похибка тоді є

$$\sigma_{\hat{p}} = \sqrt{\frac{(N-n)}{(N-1)}\cdot\frac{p(1-p)}{n}}. \qquad (2.9)$$

А з (2.4) випливає, що незміщеною оцінкою цієї дисперсії буде статистика

$$\upsilon(\hat{p}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{(N-n)}{N}\cdot\frac{\hat{p}(1-\hat{p})}{n-1}.$$

$1-\alpha$-надійний інтервал для пропорції сукупності p буде таким $\left(\hat{p}-t_\alpha\sqrt{\upsilon(\hat{p})};\ \ \hat{p}+t_\alpha\sqrt{\upsilon(\hat{p})}\right)$, або

$$\left(\hat{p}-t_\alpha\sqrt{\frac{(N-n)}{N}\cdot\frac{\hat{p}(1-\hat{p})}{n-1}};\ \ \hat{p}+t_\alpha\sqrt{\frac{(N-n)}{N}\cdot\frac{\hat{p}(1-\hat{p})}{n-1}}\right),$$

де t_α - табличне значення розподілу Стьюдента з $n-1$ степеню свободи.

При $n>50$ t_α можна замінити на c_α (див.табл.А.2 та табл.А.3 додатка).

Приклад 2.3. Щоб оцінити пропорцію виборців на користь спірної пропозиції, за допомогою простої випадкової вибірки, було відібрано $n=1200$ виборців. З них 552 повідомили, що вони підтримують пропозицію. Необхідно оцінити пропорцію населення, яке підтримує пропозицію і побудувати 0,95-надійний інтервал для цієї пропорції. Число виборців у регіоні – 1800000.

В даному випадку $N=18\cdot10^5,\ n=1200,\ \sum_{i=1}^{n}y_i=552,$ тоді $\hat{p}=\frac{552}{1200}=0{,}46.$ Оскільки $f=\frac{n}{N}=0{,}66\cdot10^{-3},$ то $1-f\approx1,$ і $\upsilon(\hat{p})\approx\frac{0{,}2484}{1199}\approx2\cdot10^{-4};$ $c_\alpha=1{,}96.$

$0{,}46-1{,}96\sqrt{2}\cdot10^{-2}<\hat{p}<0{,}46+1{,}96\sqrt{2}\cdot10^{-2}.$

Отже, $0{,}432<\hat{p}<0{,}488.$

2.6. Обсяг вибірки для оцінювання пропорції

Припустимо, що $\hat{p}$ має нормальний розподіл, дістанемо

$$P\left\{\left|\hat{p}-p\right|\geq d\right\}=\alpha.$$

Аналогічно викладкам параграфу 2.4 і враховуючи (2.9), маємо

$$c_\alpha\sqrt{\frac{(N-n)}{(N-1)}\cdot\frac{p(1-p)}{n}}=d.$$

Розв’язок цього рівняння відносно n дає формулу

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}} \approx \frac{n_0}{1 + \frac{n_0}{N}}, \qquad (2.10)$$

де $n_0 = \frac{c_\alpha^2 p(1-p)}{d^2}$.

Зауважимо, що ця формула залежить від невідомої пропорції населення p. Якщо немає попередніх оцінок p, тоді можна використати "найгірший випадок" $p = 0{,}5$ для визначення обсягу вибірки. При цьому величина $p(1-p)$ досягає найбільшого значення і, отже, обсяг вибірки n буде максимальним.

Приклад 2.4. Геолог бажає оцінити пропорцію золота в тонкій секції скелі за допомогою взяття простої випадкової вибірки n точок і відмічаючи наявність чи відсутність мінералу. Якою великою повинна бути вибірка, щоб отримати оцінку з точністю $d = 0{,}05$ для істинної пропорції з ймовірністю $0{,}95$?

Припускаємо, що N досить велике і величиною $\frac{n_0 - 1}{N}$ можна знехтувати. Оскільки p невідоме, то покладемо у формулу (2.11) $p = 0{,}5$, дістанемо

$$n \approx n_0 = \frac{(1{,}96)^2 \, 0{,}5 \cdot 0{,}5}{(0{,}05)^2} = 384{,}16.$$

Отже, обсяг вибірки 384 або 385 повинен бути достатнім, щоб задовольнити умови задачі при будь-якій фактичній пропорції сукупності p.

2.7. Оцінювання відношення

Інколи i-та одиниця може характеризуватися двома значеннями Y_i та X_i. Наприклад, Y_i - витрати i–ї сім'ї на харчування, X_i - загальний доход сім'ї.

Величина $R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i}$ називається відношенням (ratio). У нашому прикладі R – це середня частка доходу сім'ї, яка витрачається на харчування для всього населення регіону. Очевидно, що $R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{\mu_y}{\mu_x}$, тут $\mu_y = \mu$.

Оцінкою величини R може бути статистика

$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$, де (y_i, x_i) - вибіркові значення i-ї одиниці, що потрапила у вибірку, $i = 1, 2, \ldots n$. Дана оцінка буде зміщеною. Середньоквадратична похибка $E(\hat{R} - R)^2$ і дисперсія при великих n апроксимується виразом

$$D\hat{R} \approx \frac{1-f}{n\mu_x^2} \cdot \sum_{i=1}^{N} (Y_i - RX_i)^2 / (N-1), \text{ де } f = \frac{n}{N}.$$

По вибірці дану дисперсію можна оцінити за допомогою статистики

$$\upsilon(\hat{R}) = \frac{1-f}{n\bar{x}^2} \cdot \sum_{i=1}^{n} (y_i - \hat{R}x_i)^2 / (n-1).$$

Приклад 2.5. У таблиці 2.2 наведені наступні вибіркові дані: x_1 - число членів сім'ї, x_2 - місячний доход сім'ї, y - витрати сім'ї на харчування. Оцінити такі два відношення: середні витрати на харчування на одного члена сім'ї, середня частка доходу, витрачена на харчування сім'ї.

Таблиця 2.2

Вибіркові дані про склад, місячний доход та витрати сімей

№	x_1	x_2 (грн.)	y (грн)

1	3	690	320
2	2	800	500
3	3	950	550
4	4	840	600
5	2	700	380
6	1	280	130
7	3	790	350
8	5	700	500
9	2	1000	600
10	1	350	200

$n = 10,\ \bar{y} = \frac{4130}{10} = 413$(грн.) – середні витрати сім'ї на харчування.

$\sum x_1^{(i)} = 26,\ \widehat{R}_1 = \frac{\sum y_i}{\sum x_1^{(i)}} = \frac{4130}{26} \approx 158{,}8$ (грн.) – середні витрати на харчування на одного члена сім'ї.

$\sum x_2^{(i)} = 7100,\ \widehat{R}_2 = \frac{\sum y_i}{\sum x_2^{(i)}} = \frac{4130}{7100} \approx 0{,}581.$

Тобто 58,1% доходу сім'ї в середньому витрачається на харчування.

2.8. Основні поняття систематичного відбору

Простий вибірковий вибір вимагає дуже детальної роботи у процесі селекції вибірки. Наприклад, щоб отримати просту випадкову вибірку обсягу $n = 100$, треба як мінімум 100 разів звернутися до таблиці випадкових чисел. Для того, щоб спростити процедуру існує систематичний відбір (systematic sampling). Головна ідея систематичного відбору така: припустимо, що вибірка n елементів робиться із великої сукупності. Простий спосіб отримати вибірку такий, через однаковий інтервал вибирати елементи вздовж всього списку сукупності. Наприклад, вибираємо кожен десятий елемент списку. Якщо перший елемент випадковий, то такий процес вибору і буде систематичною вибіркою.

Очевидна перевага цього методу полягає в його простоті у поєднанні з належною точністю. Він дає таку саму точність результатів, як і простий випадковий відбір. Водночас необхідно дотримуватись важливої умови: необхідно, щоб не було жодної регулярності у складанні списку сукупності, яка може стати причиною тенденційності вибору. У такій ситуації витяг для вибірки треба робити з різною періодичністю. Цей метод часто використовують, коли базою даних для складання вибірки є телефонний довідник.

Нехай сукупність містить N одиниць з номерами від 1 до N. Для отримання вибірки обсягу n спочатку випадковим чином вибираємо будь-яку одиницю з перших k одиниць сукупності (це можна зробити, використовуючи датчик випадкових чисел чи статистичну таблицю (див. Додаток , табл. А.1)). Після вибору першого елемента вибираємо кожний k-ий елемент, після попереднього. Наприклад, $k=12$ і перший елемент виберемо 8. Тоді наступні елементи будуть мати номери: $20, 32, 44, 56, \ldots$. Отже перший вибраний елемент повністю визначає вибірку. Таку вибірку будемо називати систематичною вибіркою k-го порядку. Недоліком такого відбору є те, що у випадку, коли N не кратне k, у різних вибірках k-го порядку може бути не однакова кількість елементів. Наприклад, $N=21$ і $k=4$. Тоді можливі такі номери у вибірці:

1) 1, 5, 9, 13, 17, 21;
2) 2, 6, 10, 14, 18;
3) 3, 7, 11, 15, 19;
4) 4, 8, 12,16, 20.

Як бачимо, перша вибірка містить 6 елементів, а друга, третя та четверта – 5 елементів. Тобто обсяги вибірок можуть бути різними. Отже, ймовірність появи першої вибірки $\frac{6}{21}$, а інших – $\frac{5}{21}$. Якщо за оцінку середнього сукупності вибирати середнє арифметичне такої систематичної вибірки, то ця оцінка буде зміщеною.

Щоб уникнути цього можна скористатися таким методом. Вибираємо k як найбільше ціле, що лежить біля $\frac{N}{n}$. Далі випадковим чином вибираємо будь-який

елемент від 1 до N, потім беремо кожний k-й елемент, рухаючись по колу, поки не виберемо n елементів. Наприклад, $N = 21,\ n = 5$, тоді $k = 4$. Нехай вибрано елемент із номером 13. Тоді систематична вибірка 4-порядку буде містити елементи із номерами:
13, 17, 21, 4, 8. Якщо перший елемент 19, то вибірка така: 19, 2, 6, 10,14.

2.9. Оцінювання середнього та сумарного значення сукупності

Нехай $N = n \cdot k$. Тоді сукупність можна розбити на k кластерів, у кожному з яких знаходиться n одиниць. Тоді процедура випадкового вибору систематичної вибірки k-го порядку така ж сама, якщо і процедура вибору одного із k кластерів.

Таблиця 2.3

Можливі систематичні вибірки k-го порядку

	Кластер					
	1	2	…	i	…	k
	Y_1	Y_2		Y_i		Y_k
	Y_{k+1}	Y_{k+2}		Y_{k+i}		Y_{2k}
	…	…	…	…	…	…
	$Y_{(n-1)k+1}$	$Y_{(n-1)k+2}$		$Y_{(n-1)k+i}$		Y_{nk}
Середнє	$\overline{Y}_1$	$\overline{Y}_2$		$\overline{Y}_i$		$\overline{Y}_k$

Нехай $\bar{y}_{sy}$ – середнє систематичної вибірки. Тобто $\bar{y}_{sy}$ з ймовірністю $\frac{1}{k}$ дорівнює $\overline{Y}_i$, $i = \overline{1,k}$, $\hat{\tau}_{sy} = N\bar{y}_{sy}$ - оцінка сумарного значення сукупності, $\bar{y}_{sy}$ – незміщена оцінка μ.

Дійсно, $E\bar{y}_{sy} = \frac{1}{k}\sum_{i=1}^{k}\bar{Y}_i = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}Y_{ij} = \frac{1}{N}\sum_{i=1}^{N}Y_i = \mu$, де Y_{ij} – j-ий член i-ої систематичної вибірки, $j = \overline{1,n}$, $i = \overline{1,k}$, і $\sum_{i=1}^{k}\sum_{j=1}^{n}Y_{ij} = \sum_{i=1}^{N}Y_i$. Знайдемо тепер дисперсію $\bar{y}_{sy}$.

Теорема 2.4. [9, 13]. Дисперсія систематичної вибірки визначається формулою

$$D\bar{y}_{sy} = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_W^2, \qquad (2.11)$$

де

$$S_W^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_i\right)^2.$$

Наслідок 2.2. $D\bar{y} > D\bar{y}_{sy}$, тоді і тільки тоді, коли $S_W^2 > S^2$.

Таким чином, систематична вибірка точніша, ніж проста випадкова вибірка, коли одиниці всередині систематичної вибірки неоднорідні, і дає більшу похибку, коли одиниці однорідні.

Отримаємо еквівалентну формулу дисперсії $D\bar{y}_{sy}$.

Нехай $\rho_W = \frac{1}{(n-1)(N-1)S^2}\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{\substack{s=1,\\ s\neq j}}^{n}(Y_{ij}-\mu)(Y_{is}-\mu)$ – коефіцієнт кореляції між парами одиниць в одній і тій же систематичній вибірці.

Тоді

$$D\bar{y}_{sy} = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{Y}-\mu\right)^2 = \frac{S^2}{n}\left(\frac{N-1}{N}\right)\left(1+(n-1)\rho_W\right). \qquad (2.12)$$

Таким чином, формули (2.11) та (2.12) еквівалентні. Останній результат показує, що додатна кореляція між одиницями у страті призводить до збільшення дисперсії вибіркового середнього. Навіть мала додатна кореляція може мати великий ефект за рахунок множника $(n-1)$.

Дисперсія сумарного значення знаходиться, як завжди, за формулою $D\hat{\tau}_{sy} = N^2 D\bar{y}_{sy}$.

Вкажемо також на інші переваги систематичного відбору у порівнянні із простим випадковим відбором та стратифікованим :

1. Систематичну вибірку простіше здобувати і простіше виконувати правила відбору. Систематичний відбір особливо зручний у тих випадках, коли вже маємо списки одиниць, складених у тому чи іншому порядку, а також у тих випадках, коли мають справу з генеральною сукупністю, чисельність якої відома лиш наближено і одиниці якої з'являються поступово протягом якогось періоду (контроль якості продукції, що виробляється). Інколи можна отримати значну економію часу, навіть якщо вибірка здобувається до початку обстеження. Наприклад, якщо дані про всі одиниці сукупності занесені на карточки однакового розміру, що знаходяться в ящиках стандартної картотеки, тоді карточки можна здобувати із ящика, наприклад, через кожні 20 см.

2. Систематична вибірка розподілена по сукупності більш рівномірно, ніж випадкова. Це інколи робить систематичний відбір більш точним, ніж стратифікований випадковий .

Контрольні запитання, вправи

Контрольні запитання й завдання

1. *Дайте визначення простого випадкового відбору?*
2. *Які методи здобуття простої випадкової вибірки?*
3. *Напишіть формули для оцінки середнього та сумарного значень сукупност та формули для відповідних дисперсій цих оцінок.*
4. *У яких випадках використовуються табличні значення розподілу Стьюдента, а у яких – табличні значення нормального розподілу при побудові надійних інетервалів?*
5. *За якими формулами оцінюється обсяг вибірки?*
6. *Які особливості застосування загальних формул при оцінюванні пропорції сукупності?*

7. *У яких випадках використовується оцінювання відношення?*

8. *У чому відмінність систематичного відбору від простого випадкового?*

Вправи для самостійної роботи

1. На рис. 2.3 місцезнаходження об’єктів (наприклад, це можуть бути дерева, житло, шахти) в регіоні задані центрами символів “+”. Мета обстеження – оцінити число об’єктів в регіоні.

а) Регіон розбитий на 100 квадратів, $N = 100$. Зроблена проста випадкова вибірка обсягу $n = 10$. Вибрані одиниці заштриховані на рис.3. Використовуючи цю вибірку оцініть число об’єктів у регіоні, а також оцініть дисперсію вашої оцінки.

б) Використовуючи таблицю випадкових чисел, побудуйте просту випадкову вибірку обсягу $n = 10$ і обчисліть нові оцінки. Порівняйте їх з попередніми.

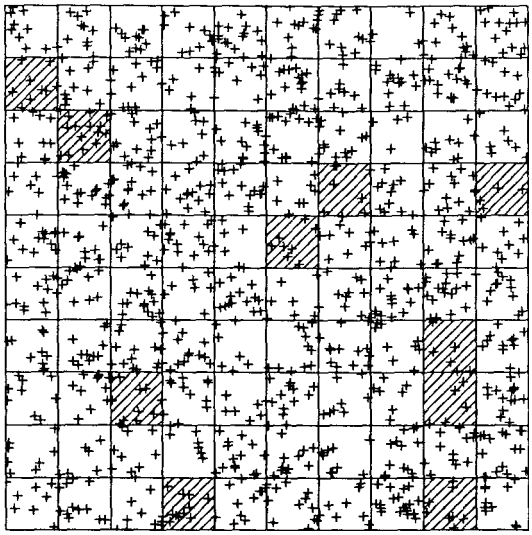

Рис. 2.3

2. Зроблена проста випадкова вибірка обсягу $n = 10$із сукупності з 200 домогосподарств. Число людей у вибраних домогосподарствах становить: 3, 4, 5, 1, 3, 3, 5, 2, 2, 3.

а) Оцінити загальне число людей в усіх домогосподарствах. Оцінити також дисперсію цієї оцінки.

б) Оцінити середнє число людей в домогосподарстві, а також оцінити дисперсії цієї оцінки.

3. Для даних частин а) та б) вправи 1 побудувати 0,95-надійний інтервал для сумарного значення сукупності. Побудувати також 0,95-надійний інтервал для середнього значення сукупності на одиницю.

4. Для даних вправи 2 побудувати 0,9-надійний інтервал:

а) для загального числа людей в домогосподарствах;

б) для середнього числа людей в одному домогосподарстві.

5. Необхідно оцінити число дерев у деякому регіоні. Область вивчення була розділена на 1000 одиниць або квадратів. З попереднього досвіду відомо, що $S^2 \approx 45$.

а) Необхідно знайти обсяг простої випадкової вибірки для оцінки загального числа дерев в регіоні з точністю до 500 дерев і надійною ймовірністю 0,95.

б) Яким буде n при точності до 1000 дерев?

6. Обробляється 3040 анкет. Проста випадкова вибірка 200 анкет при перевірці показала, що 38 анкет заповнені з помилками. Оцінити пропорцію помилкових анкет у всій сукупності, а також побудувати 0,9-надійний інтервал для цієї пропорції.

7. Яким повинен бути розмір вибірки, щоб оцінити пропорцію людей з першою групою крові серед населення у 1500 людей з рівнем точності 0,02 та надійною ймовірністю 0,95? Відносно істиного значення пропорції немає ніяких попередніх знань.

8. Дирекція приватної компанії зацікавилася в оцінці пропорції службовців, які підтримують нову інвестиційну політику. Систематична вибірка 10-го порядку була отримана, рахуючи кожного десятого, який залишив приміщення компанії у кінці робочого дня. Дані наведені у таблиці.

Номер	3	13	23	…	1993
Відгук	1	0	1		1

$N = 2000$; $\sum_{i=1}^{200} y_i = 132$.

Оцінити пропорцію та сумарне число службовців, які підтримують нову інвестиційну політику.

9. На першому курсі навчається 300 студентів. Для вивчення витрат студентів на придбання підручників була зроблена систематична вибірка 15-го порядку.

Студент	Витрати, грн.	Студент	Витрати, грн.
1	30	11	48
2	22	12	20
3	19	13	29
4	28	14	38
5	31	15	32
6	40	16	24
7	29	17	36
8	17	18	27
9	15	19	30
10	23	20	25

Оцінити середні та сумарні витрати студентів на придбання підручників.

РОЗДІЛ 3

МЕТОДИ ПРОВЕДЕННЯ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ВИБІРКОВИХ ДОСЛІДЖЕНЬ З ВИКОРИСТАННЯМ СТРАТИФІКОВАНОГО ВИПАДКОВОГО ВІДБОРУ

3.1. *Основні позначення для стратифікованого відбор.*

3.2. *Оцінювання середнього та сумарного значень сукупності.*

3.3. *Надійні інтервали.*

3.4. *Оптимальне розміщення при стратифікованому відбор.*

3.5. *Стратифікований відбір для оцінювання пропорцій.*

Дуже важливо знати, як правильно отримати вибірку і як зробити за її даними обґрунтовані висновки. Ці проблеми не грали б особливої ролі, коли матеріал, з якого проводиться відбір, був би однорідним, так що будь-яка вибірка давала б приблизно однакові результати. Проте, коли матеріал, що вивчається дуже неоднорідний, як часто і трапляється, спосіб отримання вибірки набуває вирішального значення. З цієї точки зору і розглядається один із методів побудови вибірки – стратифікований випадковий відбір. При стратифікованому відборі вся генеральна сукупність поділяється на деякі менші підсукупності, кожна з яких внутрішньо однорідна. Ці підсукупності не мають спільних одиниць і надалі будемо їх називати стратами. Вибірка робиться з окремих страт певним способом. Оскільки відбір у різних стратах робиться незалежно, то для отримання дисперсій оцінок всієї сукупності можна скласти разом дисперсії оцінок по кожній страті. Отже, принцип стратифікації полягає у поділі генеральної сукупності таким чином, щоб одиниці кожної страти були максимально подібними, однорідними, що веде до зменшення дисперсії оцінки у кожній страті. Усередині кожної страти буде спостерігатися менша варіація досліджуваної ознаки через внутрішню однорідність, властиву стратам. Це, в свою чергу, буде причиною меншої загальної варіації , спостереженої у вибірці.

Тоді, хоча страти і різняться між собою, стратифікований відбір певної кількості елементів з кожної страти буде репрезентативним для всієї сукупності в цілому, і може дати виграш в точності при оцінюванні характеристик сукупності. Інколи неоднорідну сукупність вдається розділити на підсукупності, кожна з яких внутрішньо однорідна. Наприклад, при проведенні вибіркового обстеження прибутків фірм, можна утворити як найменше дві страти: великі фірми та малі фірми.

Сукупність людей можна стратифікувати за такими ознаками, як місце проживання, розмір міста, стать, а також за соціально-економічними ознаками.

Наприклад, при проведенні вибіркових досліджень для визначення середнього рівня доходів населення, Україну можна розділити на 6 страт: західні області, східні області, південні області, центральні області, північні області, м.Київ.

Якщо при формуванні вибірки у кожній окремій страті здійснюється простий випадковий відбір, то такий відбір називається стратифікованим випадковим відбором (SRS – stratified random sampling).

3.1. Основні позначення для стратифікованого відбору

Нехай сукупність із N одиниць ділиться на L підсукупностей (страт), які складаються відповідно з $N_1, N_2, ..., N_L$ одиниць так, що $N_1 + N_2 + ... + N_L = N$. Ці сукупності не мають спільних одиниць. Нехай $n_1, n_2, ..., n_L$ – обсяги простих незалежних вибірок з цих страт, $n = n_1 + ... + n_L$ загальний обсяг вибірки.

Надалі будемо користуватися такими позначеннями для k–ї страти $(k = 1,2,...,L)$:

Y_{ki} -значення або характеристика i–ї одиниці у k–й страті $i = 1,2,...,N_k$;

y_{ki} - вибіркове значення i–ї одиниці у k–й страті $i = 1,2,...,n_k$;

$W_k = \frac{N_k}{N}$ - вага страти;

$f_k = \frac{n_k}{N_k}$ - частка відбору з k–ї страти;

$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N_k} Y_{ki}$ - істинне середнє значення k–ї страти;

$\overline{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki}$ - вибіркове середнє k–ї страти;

$\tau_k = \sum_{i=1}^{N_k} Y_{ki}$ - сумарне значення k–ї страти;

$\mu = \sum_{k=1}^{L} W_k \mu_k$ - середнє значення на одиницю всієї сукупності;

$\tau = \sum_{k=1}^{L} \tau_k$ - сумарне значення всієї сукупності;

$S_k^2 = \frac{1}{(N_k - 1)}\sum_{i=1}^{N_k} (Y_{ki} - \mu_k)^2$ - дисперсія k–ї страти;

$s_k^2 = \frac{1}{(n_k - 1)}\sum_{i=1}^{n_k} (y_{ki} - \overline{y}_k)^2$ - вибіркова дисперсія у k–й страті.

3.2. Оцінювання середнього та сумарного значень сукупності

Введемо наступні позначення

$$\overline{y}_{st} = \sum_{k=1}^{L} W_k \overline{y}_k, \ \widehat{\tau}_{st} = N\overline{y}_{st} = \sum_{k=1}^{L} N_k \overline{y}_k.$$

Індекс st - традиційне позначення для стратифікованих оцінок.

Теорема 3.1 [9, 13]. Статистики $\overline{y}_{st}$ та $\widehat{\tau}_{st}$ – незміщені оцінки μ та τ відповідно.

Знайдемо тепер дисперсії цих оцінок.

$$D\overline{y}_{st} = D\left(\sum_{k=1}^{L} W_k \overline{y}_k\right) = \sum_{k=1}^{L} W_k^2 D(\overline{y}_k) = \sum_{k=1}^{L} W_k^2 \frac{S_k^2}{n_k}(1 - f_k) = \\ = \sum_{k=1}^{L} \frac{W_k^2 S_k^2}{n_k} - \sum_{k=1}^{L} \frac{W_k^2 S_k^2}{N_k}. \qquad (3.1)$$

Тут було використано, що вибірки із кожної страти незалежні, властивості дисперсії та формулу (2.2).

Очевидно, що

$$D\hat{\tau}_{st} = N^2 D\bar{y}_{st} = \sum_{k=1}^{L} N_k^2 \frac{S_k^2}{n_k}(1 - f_k) = \sum_{k=1}^{L} N_k (N_k - n_k)\frac{S_k^2}{n_k} \qquad (3.2)$$

Стандартні похибки відповідно дорівнюють

$$\sigma_{\bar{y}_{st}} = \sqrt{\sum_{k=1}^{L} W_k^2 \frac{S_k^2}{n_k}(1 - f_k)} \quad , \sigma_{\hat{\tau}_{st}} = \sqrt{\sum_{k=1}^{L} N_k (N_k - n_k)\frac{S_k^2}{n_k}} \quad .$$

Незміщені оцінки дисперсій $D\bar{y}_{st}$ та $D\hat{\tau}_{st}$ визначаються за формулами

$$\upsilon(\bar{y}_{st}) = \sum_{k=1}^{L} W_k^2 \frac{s_k^2}{n_k}(1 - f_k) = \sum_{k=1}^{L} \frac{W_k^2 s_k^2}{n_k} - \sum_{k=1}^{L} \frac{W_k^2 s_k^2}{N_k} \qquad (3.3)$$

$$\upsilon(\hat{\tau}_{st}) = \sum_{k=1}^{L} N_k (N_k - n_k)\frac{s_k^2}{n_k} \qquad (3.4)$$

Якщо $\frac{n_k}{N_k} = \frac{n}{N}$ для всіх $k = 1,2,...,L$, то назвемо таке розміщення пропорційним.

Тоді $n_k = nW_k$ і $D(\bar{y}_{st}) = \frac{1-f}{n}\sum_{k=1}^{L} W_k S_k^2$, де $f = \frac{n}{N}$

$$D(\hat{\tau}_{st}) = \frac{(N-n)}{n}\sum_{k=1}^{L} N_k S_k^2 .$$

При пропорційному розміщенні з кожної страти роблять прості випадкові вибірки таким чином, щоб вони були пропорційні до розмірів страт у генеральній сукупності. Критерій розподілу на страти треба вибрати таким чином, щоб досліджувані одиниці були однорідні всередині кожної страти з точки зору досліджуваної змінної і відрізнялись від одиниць інших страт.

У маркетингових дослідженнях промислових ринків розмір підприємства часто використовується як критерій розподілу генеральної сукупності на страти: малі, середні і великі підприємства звичайно відмінні за свою поведінкою. У дослідженнях масового споживача класичним критерієм стратифікації є рівень прибутків.

Приклад 3.1. Оцінка потенціалу ринку тютюнових виробів України.

Для маркетолога ринок – це сукупність всіх покупців певного товару чи послуги, як реально існуючих, так і потенційних. Обсяг ринку визначається кількістю людей, здатних відреагувати на певну пропозицію. Для будь-якого виробника оцінка ринкового потенціалу є надзвичайно важливою проблемою. Адже перш, ніж збільшити обсяг товару, що виробляється, або перш, ніж вивести новий товар на ринок, або навіть перш, ніж створювати товар, який би задовольняє певну потребу, необхідно знати, яким попитом буде він користуватися. Для того, щоб прийняти рішення, на який ринок спрямовувати свої дії, фірми і оцінюють ринковий попит на товари і послуги, які вони збираються виробляти. Прогнозування попиту полягає в оцінці майбутнього попиту на основі припущень про найбільш ймовірну поведінку покупців при збереженні ряду умов у майбутньому. Для прогнозування збуту своїх товарів компанії застосовують ряд спеціальних методів, які базуються на дослідженні інформації про те, що люди мають намір купити, що вони купують зараз або що купували раніше.

Об'єктом дослідження був ринок тютюнових виробів України. Для оцінки його потенціалу був обраний метод, який, на нашу думку, є найбільш точним, оскільки базується на дослідженні інформації про наміри споживачів, тобто відображає їх реальні уподобання. Дослідження проводилось у вересні 2001 року. Інформація збиралася шляхом прямого опитування. Все населення України було поділене на страти, які співпадають з її адміністративними одиницями. У кожній страті було виділено по 2 групи: жіноче і чоловіче населення, тобто жіноче і чоловіче населення досліджувалося окремо.

У кожній групі методом простого випадкового відбору було опитано 0,1% населення. Кожна відповідь містила інформацію про найулюбленішу марку і кількість сигарет, яку людина планує випалювати протягом місяця.

Таким чином у нас була можливість застосувати точкові та інтервальні оцінки середнього та сумарного значення сукупності для стратифікованого відбору. Дані, необхідні для підрахунку точкових та інтервальних оцінок середнього та сумарного значення сукупності, представлені у Таблиці 3.1.

Спочатку оцінювалась кількість людей, що палять, по кожній групі. За всіма, відповідями людей, які палять, оцінювався потенціал тютюнового ринку України, використовуючи інформацію про кількість випалюваних протягом місяця сигарет. А потім за відповідями людей, які палять, оцінювався потенціал тютюнового ринку України щодо кожного бренду, використовуючи інформацію про найулюбленішу марку сигарет і кількість випалюваних протягом місяця сигарет. В результаті чого отримуємо, що в середньому на місяць обсяг ринку складає 6 800 303 680 сигарет, причому пересічний курець викурює в середньому 557 сигарет на місяць. Оскільки середня дисперсія в середині груп (1164,0065) значно менша міжгрупової дисперсії (6005,195), то сукупності в середині груп можна вважати відносно однорідними. Таким чином стратифікацію можна вважати вдалою, вибірку – репрезентативною, а оцінки – достатньо точними.

Таблиця 3.1

Вибіркові дані про курців у регіонах України

Регіон	Кількість курців		Кількість сигарет на місяць			
			Вибіркові середні		Вибіркові стандартні відхилення	
	Чоловіки	Жінки	Чоловіки	Жінки	Чоловіки	Жінки
Автономна Республіка Крим	326000	188000	552.69	538.54	32.78	42.3
Вінницька	281000	167000	651.76	559.11	44.22	15.43
Волинська	168000	95000	487.85	502.74	12.79	17.77
Дніпропетровська	569000	336000	567.23	556.86	33.68	29.74
Донецька	777000	448000	554.72	527.79	40.45	29.38
Житомирська	226000	126000	532.81	560.08	27.69	34.36
Закарпатська	204000	113000	513.26	506.52	31.57	34.31
Запорізька	310000	179000	513.87	560.09	39.11	21.35
Івано-Франківська	230000	131000	530.73	592.3	35.4	30.92
Київська	284000	164000	696.88	485.22	13.49	12.04
Кіровоградська	180000	104000	638	397.55	40.35	25.66
Луганська	409000	238000	635.7	624.51	56.82	34.67
Львівська	437000	240000	654.15	586.77	32.9	26.39
Миколаївська	200000	117000	681.68	556	28.41	41.7
Одеська	395000	227000	675.34	336.28	29.76	46.5
Полтавська	258000	152000	544.24	395.46	36.61	22.99
Рівненська	190000	106000	532.21	498.31	41.19	26.5
Сумська	206000	122000	561.01	396.34	23.15	28.74
Тернопільська	178000	103000	567	350.76	45.73	24.25
Харківська	458000	269000	647.37	437.73	39.5	15.33
Херсонська	189000	111000	512.53	588.75	34.44	20.78
Хмельницька	224000	131000	663.14	550.57	37.04	18.45
Черкаська	223000	132000	582.83	645.57	30.64	28.26
Чернівецька	146000	83000	474.69	452.28	37.01	37.87
Чернігівська	195000	118000	542.84	438.82	24.29	24.36
м. Київ	411000	236000	506.27	520.86	43.58	31.74
м. Севастополь	62000	35000	441.73	544.92	15.66	32.51
Всього	7736000	4471000				

Оцінки потенціалу тютюнового ринку України щодо кожного марки проводилась за відповідями людей, що віддають перевагу тій чи іншій марці сигарет. Результати цієї оцінки представлені у Таблиці 3.2.

Таблиця 3.2

Оцінки потенціалу тютюнового ринку України

Бренд	Кількість курців	Потенційний обсяг ринку
Priluky Osblyvi FF	1367000	764519580
Kozak	868000	479818870
Prima Soft	786000	437254120
Prima	591000	333905190
Priluky Osblyvi lights.	566000	321161590
L&M	497000	265689080
Prima Optima	478000	268311240
L&M Lights	475000	273316920
Bond Street Box Lights	430000	241821070
Priluky Filter	429000	243995220
Bond Street Box FF	385000	216523630
Prima Optima Lts	378000	208720190
Express 22	322000	180933830
Vatra FF KS Box	319000	178027360
Monte Carlo Box FF	262000	140897070
Marlboro Lights	209000	115455000
Vatra KS Soft	193000	106160550
Kosmos LS	185000	104968120
Chesterfield FF	181000	103516850
Otaman Filter	182000	102476040
Monte Carlo Box Lights	179000	96545450
Vatra	177000	99192620
Vatra	145000	83088690
Chesterfield Lights	131000	73694350
Slawutisch	118000	65434100
Polyot	117000	64007530
Prima Lux FF	117000	65254830
Magna FF	108000	60370290
Prima Lux Lights	107000	59835620
Magna Classic FF	99000	54346940
Marlboro	95000	52055240
Prima	86000	47293990
Astor	84000	46773620
Prima Lux Super Lights	80000	44795820
Vatra Lights KS Box	80000	43460930
Vatra Premium KS	77000	44428340
Bond Street Soft	68000	37717170
Priluky oval	64000	36184670
Berkut filter Soft	64000	36762310
West SPF M	64000	35890890

Отримані результати, близькі до даних про продаж сигарет в Україні за 2001 рік. Це свідчить про приблизну насиченість ринку.

3.3. Надійні інтервали

Коли всі обсяги вибірок із страт достатньо великі, то наближений $1-\alpha$-надійний інтервал для середньої сукупності є

$$\left(\bar{y}_{st} - c_\alpha \sqrt{\upsilon(\bar{y}_{st})};\ \bar{y}_{st} + c_\alpha \sqrt{\upsilon(\bar{y}_{st})}\ \right),$$

а для сумарного значення сукупності

$$\left(N\bar{y}_{st} - c_\alpha \sqrt{\upsilon(\hat{\tau}_{st})};\ N\bar{y}_{st} + c_\alpha \sqrt{\upsilon(\hat{\tau}_{st})}\right).$$

Як правило, нормальне наближення може використовуватися, якщо всі обсяги вибірок n_k є не меншими 30.

При малих обсягах вибірок використовується наближення з використанням розподілу Стьюдента (t- статистики). Тобто, c_α треба замінити на t_α. Число степенів свободи m підраховується за формулою [35, с.106]

$$m = \left(\sum_{k=1}^{L} a_k \cdot s_k^2\right)^2 \Bigg/ \left(\sum_{k=1}^{L} a_k{}^2 \cdot s_k^4 \Big/ (n_k - 1)\right), \qquad (3.5)$$

де $a_k = N_k(N_k - n_k)/n_k$. Якщо розміри страти N_k однакові і всі обсяги вибірок n_k також однакові, то $m = n - L$. Зауважимо, що у формулі (3.5)

$$\min_{1\le k\le L}(n_k - 1) \le m \le n - L.$$

Приклад 3.2. Побудувати 0,95-надійний інтервал для середнього сукупності, дані про яку наведені у таблиці 3.3.

Таблиця 3.3

Дані про страти

Страти	N_k	n_k	$\bar{y}_k$	s_k^2
1	20	5	1,6	3,3
2	9	3	2,8	4,0
3	12	4	0,6	2,2

$$\bar{y}_{st} = \frac{1}{41}(20\cdot 1{,}6 + 9\cdot 2{,}8 + 12\cdot 0{,}6) = 1{,}57;$$

$$\upsilon(\bar{y}_{st}) = \frac{1}{41^2}\left[20(20-5)\frac{3{,}3}{5} + 9\cdot(9-3)\frac{4{,}0}{3} + 12(12-4)\cdot\frac{2{,}2}{4}\right] = \frac{322{,}8}{1681} = 0{,}192,$$

$\widehat{\tau} = 322{,}8.\ a_1 = 60,\ a_2 = 18,\ a_3 = 24,\ m = \dfrac{322{,}8^2}{13322{,}28} = 7{,}82 \approx 8, \qquad t_\alpha = 2{,}31.$

$$1{,}57 - 2{,}31\sqrt{0{,}192} < \mu < 1{,}57 + 2{,}31\sqrt{0{,}192},$$

$$0{,}56 < \mu < 2{,}58.$$

3.4. Оптимальне розміщення при стратифікованому відборі

Нехай функція витрат на проведення вибіркового обстеження для оцінки середнього значення сукупності має лінійний вигляд $C = \sum_{k=1}^{L} c_k n_k$, де c_k - витрати у розрахунку на одну одиницю. Необхідно знайти такі оптимальні плани проведення вибіркових досліджень у кожній страті $n_1, n_2, ..., n_L$, щоб диспресія оцінки $D\bar{y}_{st}$ була мінімальною при фіксованих витратах або були мінімальні витрати при фіксованій точності оцінки.

Теорема 3.2. [9, 13]. При стратифікованому відборі з лінійною функцією витрат, дисперсія оцінки $D\bar{y}_{st}$ мінімальна при фіксованих витратах C чи витрати мінімальні при фіксованій точності V, коли

$$\frac{n_k}{n} = \frac{W_k S_k / \sqrt{c_k}}{\sum_{j=1}^{L} W_j S_j / \sqrt{c_j}} = \frac{N_k S_k / \sqrt{c_k}}{\sum_{j=1}^{L} N_j S_j / \sqrt{c_j}}\ . \qquad (3.6)$$

Отже, з цієї теореми випливає таке правило. З даної страти треба брати вибірку більшого обсягу, якщо:

1) страта більша;
2) в страті більша дисперсія;
3) відбір у страті робиться дешевше.

Зауваження 3.1. Якщо всі вартосні коефіцієнти однакові, тобто якщо $c_1 = c_2 = ... = c_L$, то $n_k = n \cdot \dfrac{W_k S_k}{\sum_{j=1}^{L} W_j S_j} = n \cdot \dfrac{N_k S_k}{\sum_{j=1}^{L} N_j S_j}$.

Цей результат відомий як нейманівське розміщення (отримано Дж.Нейманом у 1934 році [26]).

Якщо фіксовані витрати, то

$$n = \frac{C\sum_{j=1}^{L}\left(N_j S_j / \sqrt{c_k}\right)}{\sum_{k=1}^{L} N_k S_k \sqrt{c_k}}. \qquad (3.7)$$

Якщо фіксована дисперсія або точність, то

$$n = \frac{\sum_{k=1}^{L} W_k S_k \sqrt{c_k} \cdot \sum_{j=1}^{L} W_j S_j / \sqrt{c_j}}{V + \frac{1}{N}\sum_{k=1}^{L} W_k S_k^2} \qquad (3.8)$$

У випадку нейманівського розміщення досить зручний вигляд має мінімальна дисперсія

$$V_{opt} = \min D\bar{y}_{st} = \frac{\left(\sum_{k=1}^{L} W_k S_k\right)^2}{n} - \frac{\sum_{k=1}^{L} W_k S_k^2}{N}. \qquad (3.9)$$

Аргументи на користь нейманівського розміщення перед пропорційним:

- варіація досліджуваної змінної дуже відрізняється в кожній статі;
- деякі страти мають стратегічне значення істотно більше, ніж їх питома вага в генеральній сукупності;
- витрати на вибірку з одних страт можуть бути значно нижчими, ніж з інших.

Якщо порівнювати стратифікований та систематичний відбори, то систематична вибірка має приблизно таку ж точність, що і стратифікована із однією вибраною одиницею у кожній страті. Відмінність лише у тому, що при систематичному відборі в кожній страті одиниця стоїть на одному і тому ж місці, а при стратифікованому відборі воно визначається випадковим чином.

Тобто при систематичному відборі відбувається розшарування сукупності на n страт (груп), які складаються із перших k одиниць, других k одиниць і т.д.

3.5. Стратифікований відбір для оцінювання пропорцій

Як і у параграфі 2.5, пропорція – це середнє значення величин, які приймають два значення 0 і 1. Тоді у випадку стратифікованого відбору можна використати результати параграфа 3.2 з

$$\mu_k = p_k,\ \bar{y}_k = \hat{p}_k,\ S_k^2 = \frac{N_k}{N_k - 1} p_k(1 - p_k),\ s_k^2 = \frac{n_k}{n_k - 1} \hat{p}_k(1 - \hat{p}_k).$$

Отже, p_k - попорція одиниць в k-й страті,

Як і у параграфі 2.5, пропорція $\hat{p}_k$ - оцінка цієї пропорції, $p = \sum_{k=1}^{L} W_k p_k$ - пропорція одиниць в сукупності. Як наслідок теореми 3.1 незміщеною оцінкою для p буде статистика $\hat{p}_{st} = \sum_{k=1}^{L} W_k \hat{p}_k = \sum_{k=1}^{L} \frac{N_k}{N} \hat{p}_k$.

А з формули (3.1) випливає, що

$$D\hat{p}_{st} = \sum_{k=1}^{L} W_k^{\,2} \frac{(N_k - n_k)}{N_k - 1} \frac{p_k(1 - p_k)}{n_k}, \qquad (3.10)$$

а з формули (3.4) – незміщеною оцінкою цієї дисперсії буде статистика

$$\upsilon(\hat{p}_{st}) = \sum_{k=1}^{L} W_k^{\,2}(1 - f_k) \frac{\hat{p}_k(1 - \hat{p}_k)}{n_k - 1}.$$

При пропорційному розміщенні

$$D\hat{p}_{st} = \frac{(1 - f)}{nN} \sum_{k=1}^{L} N_k^2 \frac{p_k(1 - p_k)}{N_k - 1},\ f = \frac{n}{N}. \qquad (3.11)$$

Якщо всі N_k досить великі, то формула має таке наближення

$$D\hat{p}_{st} \approx \sum_{k=1}^{L} W_k^2(1 - f_k) \frac{p_k(1 - p_k)}{n_k},$$

а формула (3.11) при пропорційному розміщенні –

$$D\hat{p}_{st} \approx \frac{1 - f}{n} \sum_{k=1}^{L} W_k p_k(1 - p_k),$$

Нейманівське розміщення має вигляд

$$n_k = n \cdot \frac{N_k \sqrt{\frac{N_k}{(N_k - 1)} p_k (1 - p_k)}}{\sum_{j=1}^{L} N_j \sqrt{\frac{N_j}{(N_j - 1)} p_j (1 - p_j)}} \approx n \cdot \frac{N_k \sqrt{p_k (1 - p_k)}}{\sum_{j=1}^{L} N_j \sqrt{p_j (1 - p_j)}}$$

З теореми 3.2 отримуємо, що при лінійній функції витрат $D\hat{p}_{st}$ мінімальне при фіксованих витратах або навпаки витрати мінімальні при фіксованому рівні дисперсії, якщо

$$n_k = n \cdot \frac{N_k \sqrt{\frac{N_k}{(N_k - 1)} \frac{p_k (1 - p_k)}{c_k}}}{\sum_{j=1}^{L} N_j \sqrt{\frac{N_j}{(N_j - 1)} \frac{p_j (1 - p_j)}{c_j}}} \approx n \cdot \frac{N_k \sqrt{\frac{p_k (1 - p_k)}{c_k}}}{\sum_{j=1}^{L} N_j \sqrt{\frac{p_j (1 - p_j)}{c_j}}}.$$

Зауваження 3.2. Нехай всі c_k однакові, $k = \overline{1, L}$.

а) Стратифікований відбір для оцінювання пропорцій дає невеликий виграш у точності в порівнянні із простим випадковим відбором, якщо тільки p_k не міняються сильно від страти до страти.

б) Оптимальне розміщення при заданому n дає невеликий виграш в точності в порівнянні із пропорційним розміщенням, якщо для всіх страт $p_k \in [0{,}1;\ 0{,}9], k = \overline{1, L}$.

Приклад 3.3. Нехай $L = 3,\ W_1 = W_2 = W_3 = 1/3$, N та всі N_k досить великі.

Розглянемо просту випадкову вибірку та пропорційний стратифікований відбір.

$$D\hat{p} \approx \frac{1 - f}{n} p(1 - p);\ D(\hat{p}_{st}) \approx \frac{1 - f}{n} \sum_{k=1}^{3} W_k p_k (1 - p_k).$$

$z = \frac{D(\hat{p})}{D(\hat{p}_{st})} \cdot 100\% \approx \frac{p(1 - p)}{\frac{1}{3} \sum_{k=1}^{3} p_k (1 - p_k)} \cdot 100\%$ - відносний виграш в точності при стратифікованому відборі.

а) Нехай $p_1 = 0{,}4;\ p_2 = 0{,}5;\ p_3 = 0{,}6;\ p = \sum_{k=1}^{3} W_k p_k = 0{,}5;$

$$z = \frac{0{,}25}{0{,}2433} \cdot 100\% = 103\%.$$

б) $p_1 = 0{,}3; p_2 = 0{,}5; p_3 = 0{,}7; p = 0{,}5;$

$$z = \frac{0{,}25}{0{,}2233} \cdot 100\% = 112\%.$$

в) $p_1 = 0{,}2; p_2 = 0{,}5; p_3 = 0{,}8; p = 0{,}5;$

$$z = \frac{0{,}25}{0{,}19} \cdot 100\% = 132\%.$$

г) $p_1 = 0{,}1; p_2 = 0{,}5; p_3 = 0{,}9; p = 0{,}5;$

$$z = \frac{0{,}25}{0{,}1433} \cdot 100\% = 174\%.$$

З цього прикладу видно, що у випадку, коли страти неоднорідні, виграш в точності при стратифікованому відборі досить великий.

Контрольні запитання, вправи

Контрольні запитання й завдання

1. *Дайте визначення стратифікованого випадкового відбору.*
2. *Коли доцільно використовувати стратифікований відбір?*
3. *Які статистики використовуються для оцінювання середнього та сумарного значень сукупності?*
4. *Дайте визначення пропорційного та нейманівського розміщення.*
5. *Які аргументи на користь нейманівського розміщення перед пропорційним?*

Вправи для самостійної роботи

1.

Страти	N_i	n_i	$\overline{y}_i$	s_i^2
1	100	50	10	280
2	50	25	20	70

3	300	50	30	60

а) Оцінити середнє сукупності.

б) Побудувати 0,95-надійний інтервал для середнього сукупності.

2.Нехай $n = 100,\ N_1 = 200,\ N_2 = 300,\ S_1^2 = 81;\ S_2^2 = 16$.

а) Обчислити пропорційне розміщення.

б) Обчислити нейманівське розміщення.

в) Обчислити оптимальне розміщення, якщо $c_1 = 4$ грн.; $c_2 = 9$ грн.

3. Досліджується громадська думка про проблеми міста. 90% всіх родин у місті мають телефон. Вартість телефонного інтерв'ю родини складає 1 грн., а вартість персонального інтервью родини без телефонування складає 4 грн.(включаючи витрати на проїзд). На опитування виділено 5000 грн. Припустимо, що дисперсії в телефонній і нетелефонній стратах однакові. Знайти оптимальні розміри вибірок в кожній страті, тобто у скількох родин повинні брати інтерв'ю в кожній страті?

4. У місті вивчаються ціни на двокімнатні квартири з метою оцінки середньої ціни. Всю сукупність квартир, що продаються, можна розбити на 2 страти: з високою ціною, та з низькою ціною. В середньому ціна на квартири з першої страти в 4 рази вища, ніж ціна квартири з другої страти, тобто $\mu_1 = 4\mu_2$.Вважати, що S_k пропорційне $\sqrt{\mu_k}$.Відомо, що $N_1 = 400,\ N_2 = 2000$. Як розподілити вибірку обсягом $n = 100$ між двома стратами?

РОЗДІЛ 4

МЕТОДИ ПРОВЕДЕННЯ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ВИБІРКОВИХ ДОСЛІДЖЕНЬ З ВИКОРИСТАННЯМ КЛАСТЕРНОГО РІВНОІМОВІРНІСНОГО ВІДБОРУ

4.1. *Основні позначення для кластерного відбору.*

4.2. *Одноступінчастий кластерний відбір.*

4.3. *Двоступінчастий кластерний відбір.*

4.4 *Оптимальне розміщення при двоступінчастому кластерному відборі.*

4.4. *Двоступінчастий стратифікований відбір.*

Нагадаємо, що метою вибіркових досліджень є визначення необхідної інформації про параметри сукупності з мінімальними витратами. Стратифікований відбір часто дає точніші оцінки, ніж простий випадковий відбір. Систематичний відбір часто дає не менш точні оцінки, як і простий випадковий відбір, але простіший у виконанні. У цьому розділі ми розглянемо новий тип відбору – кластерний.

Кластерна вибірка – це імовірнісна вибірка, у якій кожна вибіркова одиниця є групою або кластером елементів (cluster sampling). Кластерний відбір може бути дешевшим, ніж простий випадковий чи стратифікований відбір, якщо вартість отримання всього списку сукупності висока чи якщо вартість обстеження всередині одиниць значно менша, ніж вартість переїзду між одиницями.

Для ілюстрації цього твердження припустимо, що необхідно оцінити середні прибутки на сім’ю у великому місті. Як треба побудувати вибірку? Якщо ми будемо використовувати просту випадкову вибірку, то нам необхідно мати повний список сімей у місті, а це може бути дуже дорого чи неможливо отримати. Ми також не зможемо використовувати стратифіковану вибірку тому, що знову таки необхідно знати скільки сімей у кожній страті. Простіше скористатися випадковою вибіркою сімей, поділивши місто на райони чи блоки (кластери сімей) за допомогою простої

випадкової вибірки вибрати блоки з популяції. А далі у кожному блоці чи кластері провести опитування всіх сімей.

Для ілюстрації другого твердження нехай $N = 10\,000$ сімей. Нам потрібно зробити вибіркове опитування $n = 400$ сімей. Це може бути проста випадкова вибірка. Тоді основні витрати будуть на переїзд від однієї сім'ї до іншої по всьому місту. А можна розбити всю сукупність, скажімо, на 500 кластерів по 20 сімей у кожному. Причому сім'ї з одного кластеру живуть поруч. Тоді нам достатньо випадковим чином вибрати 20 кластерів і всередині кожного кластера провести опитування. Основні витрати будуть тепер лише на переїзд між 20 кластерами.

Перевага такого способу отримання вибірки полягає в тому, що він простий і економний. Немає необхідності, наприклад, складати списки всіх сімей у місті, їх адреси, досить визначити адреси кластерів. Недоліки цього методу полягають у двох джерелах можливих похибок отримання вибірки: одне з них пов'язане із відбором кластерів, а інше – з відбором сімей.

Щоб процедура відбору була ефективною, окремі кластери повинні бути "досліджуваною сукупністю в мініатюрі", репрезентативною щодо генеральної сукупності. Тому кожен кластер повинен бути гетерогенним (йому має бути притаманна вся різноманітність властивостей) і подібним до решти кластерів (майже ідентичним) . За такого способу складання вибірки часто використовують телефонні довідники, у яких сторінку розглядають як окремий блок. Якщо в довіднику 400 сторінок і 100 прізвищ на сторінці, то, щоб скласти вибірку з 300 спостережень, досить випадково вибрати 3 сторінки і включити у вибірку всі прізвища, розміщені на них. Легко зрозуміти, що для того, щоб ця процедура була ефективною, алфавітний порядок прізвищ ніяк не повинен впливати на результати опитування. Однак ця гіпотеза не завжди правильна.

Що, наприклад, відбудеться з вибіркою з багатонаціональної сукупності людей, якщо одна з трьох сторінок здебільшого містить прізвища, характерні для певної національності?

Чим же відрізняється кластерний від стратифікованого відбору? Нехай кластери співпадають зі стратами. Тоді кластерний відбір відрізняється від

стратифікованого тим, що при стратифікованому відборі вибираються одиниці з кожної страти, а при кластерному відборі вибираються лише кластери і повністю обстежуються всі його одиниці.

Зрозуміло, що дисперсія оцінки при стратифікованому відборі буде залежати від варіації значень всередині страти. А при кластерному відборі дисперсія оцінка буде головним чином залежати від варіації між кластерами.

Виділимо основі переваги кластерного відбору у порівнянні з іншими видами відбору :

- переваги організаційного характеру (обстежувана генеральна сукупність складається із розрізнених груп одиниць або серій);
- можна використовувати у тих випадках, коли достовірного списку елементів сукупності не існує або його складання вимагає значних витрат (списки населення, будинків, ферм), але при цьому можуть бути виділені природні відокремлені групи одиниць – квартали у містах, участки землі і таке інше;
- широко використовується при статистичному контролі якості готової продукції, наприклад, при використанні „мірної" тари, коли деталі одного типового розміру складають у стандартні ящики;
- на практиці найчастіше зустрічається кластер ний відбір з рівновеликими (однакового розміру) кластерами, що спрощує розрахунки.

До недоліків кластерної вибірки можна віднести відносну однорідність одиниць всередині кластеру, можливість обстеження лише невеликого числа кластерів, досить високий рівень похибок вибірки.

4.1. Основні позначення для кластерного відбору

При простому випадковому відборі одиниці, які вибираються є також елементами спостереження. При кластерному відборі, вибірковими одиницями є кластери, а елементи, які спостерігаються є внутрішніми у кластері.

Нехай сукупність складається із N кластерів або одиниць,

M_i - число елементів у i-ій одиниці,

$K = \sum_{i=1}^{N} M_i$ - загальне число елементів у сукупності.

Y_{ij} – значення j-го елемента i-ї одиниці, $i = \overline{1, N}$, $j = \overline{1, M_i}$;

$\tau_i = \sum_{j=1}^{M_i} Y_{ij}$ – сумарне значення елементів у i-й одиниці;

$\tau = \sum_{i=1}^{N} \tau_i$ - сумарне значення сукупності;

$\mu = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} \cdot \frac{1}{K} = \frac{\tau}{K}$ – середнє значення елементів сукупності;

$\overline{Y_i} = \frac{\tau_i}{M_i} = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$ – середнє значення елементів у i-й одиниці;

$S_i^2 = \frac{1}{(M_i - 1)} \sum_{j=1}^{M_i} \left(Y_{ij} - \overline{Y_i}\right)^2$ – дисперсія елементів i-ї одиниці;

$S^2 = \frac{1}{(K-1)} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(Y_{ij} - \mu\right)^2$ – дисперсія всієї сукупності;

$S_\tau^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} \left(\tau_i - \frac{\tau}{N}\right)^2$ – дисперсія між сумарними значеннями.

Позначення для вибірки:

n – число одиниць у вибірці;

m_i – число елементів у вибірці із i-ї одиниці, $i = \overline{1, n}$;

y_{ij} – вибіркове значення j-го елемента i-ї одиниці, $j = \overline{1, m_i}$, $i = \overline{1, n}$;

$\overline{y_i} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ – вибіркове середнє значення елементів у i-ї одиниці;

$\widehat{\tau_i} = \sum_{j=1}^{m_i} \frac{M_i}{m_i} y_{ij}$ – оцінка сумарного значення елементів у i-ї одиниці;

$\widehat{\tau} = \sum_{i=1}^{n} \frac{N}{n} \widehat{\tau_i}$ – незміщена оцінка сумарного значення сукупності;

$\overline{\overline{y}} = \frac{\widehat{\tau}}{K}$ – незміщена оцінка середнього значення сукупності;

$s_\tau^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\tau}_i - \dfrac{\hat{\tau}}{N}\right)^2$ – незміщена оцінка дисперсії S_τ^2;

$s_i^2 = \dfrac{1}{m_i - 1}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2$ – вибіркова дисперсія у i-ї одиниці.

4.2. Одноступінчастий кластерний відбір

При одноступінчастому кластерному відборі всі елементи, які є у вибраній одиниці (кластері), підлягають обстеженню. Тобто $m_i = M_i$, $i = \overline{1,n}$.

Кластери однакового розміру

Розглянемо спочатку найпростіший випадок, коли всі кластери (одиниці) мають однаковий розмір: $M_i = m_i = M$, $K = N \cdot M$. Тоді

$$\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n}\tau_i,\ \bar{\bar{y}} = \frac{\hat{\tau}}{K} = \frac{N}{n}\cdot\sum_{i=1}^{n}\sum_{j=1}^{M}Y_{ij}\cdot\frac{1}{NM} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i.$$

Як наслідок розділу 2 (вибірка n одиниць із сукупності N одиниць) $\hat{\tau}$ і $\bar{\bar{y}}$ – незміщені оцінки τ та μ і

$$D\hat{\tau} = N^2(1-f)\frac{S_\tau^2}{n},\quad f = \frac{n}{N}, \tag{4.1}$$

$$D\bar{\bar{y}} = (1-f)\frac{S_\tau^2}{nM^2}.$$

При цьому незміщеними оцінками цих дисперсій будуть

$$\upsilon(\hat{\tau}) = N^2(1-f)\frac{s_\tau^2}{n}, \tag{4.2}$$

$$\upsilon(\bar{\bar{y}}) = (1-f)\frac{s_\tau^2}{nM^2}, \tag{4.3}$$

де $s_\tau^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(\tau_i - \dfrac{\hat{\tau}}{N}\right)^2$.

Приклад 4.1. Представник ректорату бажає визначити середні добові витрати на харчування студентів, що живуть у гуртожитку. У гуртожитку 100 кімнат, у кожній з яких живе 4 студента. Представник випадковим чином вибирає 5 кімнат і

опитує всіх студентів, які живуть у цих кімнатах. Результати у гривнях наведені у таблиці 4.1.

Таблиця 4.1.

Кластерна вибірка кімнат студентів

№ персони	Кімната (кластер)				
	1	2	3	4	5
1	3,08	2,36	2,00	3,00	2,68
2	2,60	3,04	2,56	2,88	1,92
3	3,44	3,28	2,52	3,44	3,28
4	3,04	2,68	1,88	3,64	3,20
Сума	12,16	11,36	8,96	12,96	11,08

Отже, $N=100$, $n=5$, $M=4$, $K=400$,

$$\hat{\tau}=\frac{100}{5}(12{,}16+11{,}36+8{,}96+12{,}96+11{,}08)=1130{,}4;$$

$$s_\tau^2=\frac{1}{4}\Big((12{,}16-11{,}304)^2+(11{,}36-11{,}304)^2+(8{,}96-11{,}304)^2+$$
$$+(12{,}96-11{,}304)^2+(11{,}08-11{,}304)^2\Big)=2{,}256.$$

Тоді $\bar{\bar{y}}=\frac{1130{,}4}{400}=2{,}826$ (грн.) і стандартна похибка цієї оцінки

$$SE(\bar{\bar{y}})=\sqrt{\upsilon(\bar{\bar{y}})}=\sqrt{\left(1-\frac{5}{100}\right)\frac{2{,}256}{5\cdot 4^2}}=0{,}164.$$

Нехай $1-\alpha=0{,}95$. Тоді $k=n-1=4$ – число степенів свободи; $t_\alpha=2{,}77$ і

$$2{,}826-2{,}77\cdot 0{,}164<\mu<2{,}826+2{,}77\cdot 0{,}164,$$

$$2{,}37<\mu<3{,}28.$$

Якщо $S_c^2<S^2$, то кластерний відбір має точнішу оцінку, ніж простий випадковий відбір. Тут $S_c^2=\frac{1}{(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{Y}_i-\mu)^2$ – середньоквадратичне відхилення між одиницями (кластерами), а S^2 - дисперсія всієї сукупності елементів. Справедлива також інша формула

$$S^2=\frac{(N-1)S_c^2}{NM-1}+\frac{N(M-1)}{NM-1}S_W^2,$$

де $S_W^2 = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(Y_{ij} - \bar{Y}_i\right)^2$ – середньоквадратичне відхилення всередині одиниць. Тоді якщо $S^2 > S_c^2$, то $(N-1)S_c^2 + N(M-1)S_W^2 > (NM-1)S_c^2$, або $N(M-1)S_W^2 > N(M-1)S_c^2$. Тобто $S_W^2 > S_c^2$.

Таким чином, нерівність $S^2 > S_c^2$ еквівалентна такій $S_W^2 > S_c^2$ і кластерний відбір дає точнішу оцінку ніж простий випадковий відбір, якщо середньоквадратичне відхилення між елементами всередині одиниць більше ніж середньоквадратичне відхилення середніх значень елементів між одиницями. Це означає, що між собою кластери в середньому не сильно відрізняються, але всередині кожен кластер повинен бути неоднорідним.

Приклад 4.2. Нехай маємо дві популяції, кожна з яких має три кластери з трьома елементами: $N = 3,\ M = 3,\ K = 9$.

Таблиця 4.2

Розподіл елементів по кластерам у двох популяціях

	Сукупність А			Сукупність В		
Кластер 1	6	12	18	5	6	7
Кластер 2	7	13	19	13	11	12
Кластер 3	5	11	17	17	18	19

Обидві сукупності мають однакові середні і дисперсії сукупності: $\mu = 12,\ S^2 = 27{,}75$. У сукупності А більша варіація всередині кластерів, а у сукупності В між кластерами.

Таблиця 4.3.

Середні та дисперсії по кластерам у двох популяціях

	Сукупність А		Сукупність В	
	$\bar{Y}_i$	S_j^2	$\bar{Y}_i$	S_j^2
Кластер 1	12	36	6	1
Кластер 2	13	36	12	1
Кластер 3	11	36	18	1

Отже, сукупність А має більшу варіацію між елементами всередині кластерів, але меншу варіацію між середніми у кластерах:

$S_c^2 = \frac{6}{2} = 3$, $S_W^2 = \frac{216}{6} = 36$.Елементи у одному кластері менш подібні, ніж випадково вибрані елементи із всієї сукупності.

Для сукупності В маємо протилежну ситуацію. Більша варіація між кластерами, а всередині кластери однорідні: $S_c^2 = \frac{216}{2} = 108$, $S_W^2 = \frac{6}{6} = 1$.

Отже, для сукупності В більш ефективний простий випадковий відбір, а для сукупності А – кластерний відбір.

Кластери неоднакового розміру

Нехай тепер M_i різні. Тоді незміщені оцінки τ та μ наступні:

$\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n}\tau_i$, $\bar{\bar{y}} = \frac{\hat{\tau}}{K}$, де $K = \sum_{i=1}^{N} M_i$.

$D\bar{\bar{y}}$ та $\upsilon\left(\bar{\bar{y}}\right)$ визначаються за формулами (4.1)–(4.3),

$$D\bar{\bar{y}} = N^2\left(1-\frac{n}{N}\right)\frac{S_\tau^2}{n}\cdot\frac{1}{K^2} = \frac{1}{\overline{M}^2}\left(1-\frac{n}{N}\right)\frac{S_\tau^2}{n}, \text{ де } \overline{M} = \frac{K}{N},$$

$$\upsilon\left(\bar{\bar{y}}\right) = \left(1-\frac{n}{N}\right)\frac{s_\tau^2}{n}\cdot\frac{1}{\overline{M}^2}.$$

Можна побудувати оцінки τ та μ по відношенню. Якщо всі M_i відомі, то для оцінки по відношенню M_i буде аналогом X_i (див. параграф 4.1).

$$\hat{\tau}_R = \frac{\sum_{i=1}^{n}\tau_i}{\sum_{i=1}^{n} M_i}\cdot K,\ \tau_X = K, \qquad \hat{\mu}_R = \frac{\hat{\tau}_R}{K} = \frac{\sum_{i=1}^{n}\tau_i}{\sum_{i=1}^{n} M_i} \tag{4.4}$$

Зауважимо, що $R = \frac{\tau}{K} = \mu$ – середнє значень елементів сукупності. Отже, $\hat{\mu}_R = \hat{R}$ – оцінка відношення.

Відмітимо, що для обчислення оцінки $\hat{\tau}_R$ необхідно знати $K = \sum_{i=1}^{N} M_i$ – сумарне число елементів, а для обчислення $\hat{\tau}$ – цього не потрібно. Однак, для обчислення оцінки середнього $\hat{\mu}_R$ не потрібно знати K, а для обчислення $\bar{\bar{y}}$ – це необхідно.

$$D\hat{\tau}_R \approx \frac{N^2}{n}\left(1-\frac{n}{N}\right)\cdot\frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i-\mu\right)^2\cdot M_i^2,$$

$$D\hat{\mu}_R \approx \frac{1}{n}\left(1-\frac{n}{N}\right)\cdot\frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i-\mu\right)^2\cdot\frac{M_i^2}{\bar{M}^2}.$$

Відповідні оцінки цих дисперсій

$$\upsilon\left(\hat{\tau}_R\right)=\frac{N^2}{n}\left(1-\frac{n}{N}\right)\cdot\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(\tau_i-M_i\hat{\mu}_R\right)^2,$$

$$\upsilon\left(\hat{\mu}_R\right)=\frac{1}{n\left(\bar{M}\right)^2}\left(1-\frac{n}{N}\right)\cdot\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(\tau_i-M_i\hat{\mu}_R\right)^2.$$

Дисперсія оцінки по відношенню залежить від варіації (мінливості) середніх значень елементів у кластерах і може бути набагато меншою ніж дисперсія незміщеної оцінки.

4.3. Двоступінчастий кластерний відбір

При двоступінчастому відборі вибірка береться в два етапи:

1. Вибирається n одиниць із сукупності у N одиниць (проста випадкова вибірка).
2. Вибирається m_i елементів із вибраної i-ї одиниці (кластера), $i=\overline{1,n}$ (проста випадкова вибірка обсягу m_i із сукупності в M_i одиниць).

Наприклад, при оцінці середнього величини врожаю пшениці в Україні, одиницями будуть області України, де вирощують пшеницю, а елементи – райони в області. Незміщена оцінка τ наступна

$\hat{\tau}=\frac{N}{n}\sum_{i=1}^{n}\hat{\tau}_i=\frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{M_i}{m_i}y_{ij}=\frac{N}{n}\sum_{i=1}^{n}M_i\bar{y}_i$, де $\bar{y}_i=\frac{1}{m_i}\sum_{j=1}^{m_i}y_{ij}$ – вибіркове середнє значення елемента у i-й одиниці. Дисперсія цієї оцінки така [9, 13, 25]:

$$D\hat{\tau} = N^2\left(1-\frac{n}{N}\right)\frac{S_\tau^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)\cdot M_i^2\cdot\frac{S_i^2}{m_i}, \qquad (4.5)$$

де S_τ^2 – дисперсія між сумарними значеннями одиниць, а S_i^2 – дисперсія між елементами i-ї одиниці (див. параграф 4.1). Для оцінки $D\hat{\tau}$ покладемо

$$s_\tau^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\tau}_i - \frac{\hat{\tau}}{N}\right)^2 \text{ і } s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij}-\bar{y}_i\right)^2.$$

Тоді незміщеною оцінкою $D(\hat{\tau})$ буде статистика

$$\upsilon(\hat{\tau}) = N^2\left(1-\frac{n}{N}\right)\frac{s_\tau^2}{n} + \frac{N}{n}\sum_{i=1}^{n}\left(1-\frac{m_i}{M_i}\right)\cdot M_i^2\cdot\frac{s_i^2}{m_i} \qquad (4.6)$$

Як бачимо, перший доданок у формулі (4.5) – це дисперсія одноступінчастого кластерного відбору, а другий доданок – це дисперсія другого етапу кластерного відбору.

Далі $\bar{\bar{y}} = \frac{\hat{\tau}}{K}$, і

$$D\bar{\bar{y}} = \frac{D\hat{\tau}}{K^2} = \frac{1}{\overline{M}^2}\left(1-\frac{n}{N}\right)\frac{S_\tau^2}{n} + \frac{1}{n\overline{M}^2 N}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i}, \quad \upsilon(\bar{\bar{y}}) = \frac{\upsilon(\hat{\tau})}{K^2}.$$

Наведемо альтернативні оцінки по відношенню. Як і у попередньому параграфі

$$\hat{\mu}_R = \frac{\sum_{i=1}^{n}\hat{\tau}_i}{\sum_{i=1}^{n}M_i} = \frac{\sum_{i=1}^{n}M_i\bar{y}_i}{\sum_{i=1}^{n}M_i}, \ \hat{\tau}_R = \frac{\sum_{i=1}^{n}M_i\bar{y}_i}{\sum_{i=1}^{n}M_i}K,$$

$$D\hat{\mu}_R \approx \frac{1}{\overline{M}^2}\left(\left(1-\frac{n}{N}\right)\frac{S_R^2}{n} + \frac{1}{nN}\sum_{i=1}^{N}M_i^2\cdot\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i}\right), \text{ де}$$

$$S_R^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\tau_i - M_i\mu\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}M_i^2\left(\bar{Y}_i-\mu\right)^2,$$

$D\hat{\tau}_R = K^2 D\hat{\mu}_R$ [9, 13, 25].

Оцінки цих дисперсій будуть такими

$$\upsilon(\hat{\mu}_R) = \frac{1}{\overline{M}^2}\left(\left(1-\frac{n}{N}\right)\frac{s_R^2}{n} + \frac{1}{nN}\sum_{i=1}^{N}M_i^2\cdot\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right), \text{ де}$$

$$s_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_i - \hat{\mu}_R\right)^2 M_i^2\,,\ \upsilon\left(\hat{\tau}_R\right) = \upsilon\left(\hat{\mu}_R\right)\cdot K^2\,.$$

Приклад 4.3. Взято просту випадкову вибірку обсягу $n=3$ одиниці із сукупності $N=100$ одиниць. З кожної одиниці вибираються два елемента $m_i=2$ за допомогою простої випадкової вибірки. Розміри вибраних одиниць $M_1=24$, $M_2=20$, $M_3=15$. Відомо, що $y_{11}=8$, $y_{12}=12$, $y_{31}=1$, $y_{32}=3$, $y_{21}=y_{22}=0$. Оцінити τ та μ.

$$\hat{\tau}_1 = \frac{24}{2}(8+12) = 240\,,\hat{\tau}_2 = \frac{20}{2}(0+0) = 0\,,\hat{\tau}_3 = \frac{15}{2}(1+3) = 30\,,$$

Тоді $\hat{\tau} = \frac{100}{3}(240+0+30) = 9000$.

Оскільки $K = \sum_{i=1}^{N} M_i$ нам невідомо, то скористаємося оцінкою по відношенню

$$\hat{\mu}_R = \frac{240+0+30}{24+20+15} = \frac{270}{59} \approx 4{,}58\,,$$

$\bar{\bar{y}}_1 = \frac{\hat{\tau}}{N} = 90$ – середнє значення на одну одиницю.

Тоді $s_\tau^2 = \frac{1}{3-1}\left((240-90)^2 + (0-90)^2 + (30-90)^2\right) = 17100\,,$

$\bar{y}_1 = 10;\ \bar{y}_2 = 0;\ \bar{y}_3 = 2;$

$s_1^2 = \frac{1}{2-1}\left((8-10)^2 + (12-10)^2\right) = 8;\ s_2^2 = 0;$

$s_3^2 = \frac{1}{2-1}\left((1-2)^2 + (3-2)^2\right) = 2;$

$$\upsilon(\hat{\tau}) = 100\cdot(100-9)\cdot\frac{17100}{3} +$$
$$+\frac{100}{3}\cdot\left(24\cdot(24-2)\cdot\frac{8}{2} + 20\cdot(20-2)\cdot\frac{0}{2} + 15\cdot(15-2)\cdot\frac{2}{2}\right) = 55366900;$$

$SE(\hat{\tau}) = \sqrt{\upsilon(\hat{\tau})} = 7441;$

$s_R^2 = \frac{1}{3-1}\left((10-4{,}58)^2\cdot 24^2 + (0-4{,}58)^2\cdot 20^2 + (2-4{,}58)^2\cdot 15^2\right) = 13404{,}5\,.$

Замість $\overline{M}$, яке нам невідоме, використовуємо $\overline{M}_R$:

$$\overline{M}_R = \frac{59}{3} \approx 19{,}7;$$

$$\upsilon(\hat{\mu}_R) \approx \frac{1}{(19{,}7)^2}\left(\frac{97}{100}\cdot\frac{13404{,}5}{3} + \right.$$

$$\left. + \frac{1}{300}\left(24\cdot(24-2)\cdot\frac{8}{2} + 20\cdot(20-2)\cdot\frac{0}{2} + 15\cdot(15-2)\cdot\frac{2}{2}\right)\right) = 11{,}19;$$

$$SE(\hat{\mu}_R) = \sqrt{\upsilon(\hat{\mu}_R)} = \sqrt{11{,}19} \approx 3{,}345\,.$$

4.4. Оптимальне розміщення при двоступінчастому кластерному відборі

Нехай всі $M_i = M$ і $m_i = m$. Тоді

$$D\bar{\bar{y}} = \left(1 - \frac{n}{N}\right)\frac{S_c^2}{nM} + \left(1 - \frac{m}{M}\right)\frac{S_W^2}{nm}, \text{ де}$$

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\bar{Y}_i - \mu\right)^2 = \frac{M}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \mu\right)^2,$$

$$S_W^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}_i\right)^2 \text{ (див. параграф 4.2).}$$

Нехай $C = c_1 n + c_2 nm$ – функція витрат або функції вартості обстеження, де $c_1 n$ – вартість на проведення першого етапу відбору, а $c_2 nm$ – другого етапу відбору. Тут c_1 – вартість відбору однієї одиниці, а c_2 – вартість відбору одного елемента. Подамо $D\bar{\bar{y}}$ у вигляді

$$V = D\bar{\bar{y}} = \frac{1}{nM}\left(S_c^2 - S_W^2\right) + \frac{1}{mn}S_W^2 - \frac{S_c^2}{MN}.$$

Тоді мінімізація $D\bar{\bar{y}}$ при фіксованих витратах C, шляхом вибору n та m, чи мінімізація витрат C при фіксованій точності, еквівалентна мінімізації добутку

$$\left(V + \frac{S_c^2}{MN}\right)C = \left(c_1 + c_2 m\right)\left(\frac{S_c^2 - S_W^2}{M} + \frac{S_W^2}{m}\right).$$

Використовуючи нерівність Коші–Буняковського дістанемо, що

$$m_{opt} = \frac{S_W \sqrt{M}}{\sqrt{S_c^2 - S_W^2}} \cdot \sqrt{\frac{c_1}{c_2}} = \frac{S_w}{S_u} \sqrt{\frac{c_1}{c_2}}.$$

Припускається, що $S_c^2 > S_W^2$. Надалі будемо позначати $S_u^2 = \frac{S_c^2 - S_W^2}{M}$.

Округляємо m_{opt} до найближчого цілого. Якщо m таке ціле, що $m < m_{opt} < m+1$, то вибираємо $m+1$ у випадку $m_{opt}^2 \geq m(m+1)$, і m, коли $m_{opt}^2 < m(m+1)$. Коли ж $m_{opt} > M$ чи $S_c^2 \leq S_W^2$, то $m = M$ і використовуємо одноступінчастий відбір. Якщо фіксовані витрати, то $n_{opt} = \frac{C}{c_1 + c_2 m_{opt}}$.

Якщо фіксована дисперсія V, то

$$n_{opt} = \frac{S_u^2 + \frac{S_W^2}{m_{opt}}}{V + \frac{S_c^2}{MN}} = \frac{S_u \left(S_u + S_W \sqrt{\frac{c_2}{c_1}} \right)}{V + \frac{S_c^2}{MN}}.$$

4.5. Двоступінчастий стратифікований відбір

Нехай тепер сукупність, яка вивчається, ділиться на L страт, k-та страта складається з N_k основних одиниць, кожна з яких містить M_k елементів. Нехай $n_1, n_2, ..., n_L$ - обсяги простих незалежних випадкових вибірок одиниць з цих страт, а $m_1, m_2, ..., m_L$ – обсяги простих незалежних випадкових вибірок елементів з кожної вибраної одиниці відповідної страти. Будемо користуватися такими позначеннями:

$Y_{ij}^{(k)}$ - значення j-го елемента i-ї одиниці k-ї страти; $i = \overline{1, N_k}$, $j = \overline{1, M_k}$, $k = \overline{1, L}$;

$y_{ij}^{(k)}$ - вибіркове значення j-го елемента i-ї одиниці k-ї страти; $i = \overline{1, n_k}$, $j = \overline{1, m_k}$, $k = \overline{1, L}$;

$\overline{y}_i^{(k)} = \sum_{j=1}^{m_k} y_{ij}^{(k)} / m_k$ - вибіркове середнє значення елемента i-ї одиниці k-ї страти;

$\overline{Y}_i^{(k)} = \sum_{j=1}^{M_k} Y_{ij}^{(k)} / M_k$ - істинне середнє значення елемента i-ї одиниці k-ї страти;

$\bar{y}^{(k)} = \sum_{i=1}^{n_k} \bar{y}_i^{(k)} / n_k$ - вибіркове середнє значення елемента k-ї страти;

$\mu^{(k)} = \sum_{i=1}^{N_k} \bar{Y}_i^{(k)} / N_k$ - істинне середнє k-ї страти;

$S_{1k}^2 = \sum_{i=1}^{N_k} \left(\bar{Y}_i^{(k)} - \mu^{(k)} \right)^2 / \left(N_k - 1 \right)$ - істинна дисперсія між середніми значеннями одиниць k-ї страти;

$S_{2k}^2 = \sum_{i=1}^{N_k} \sum_{j=1}^{M_k} (Y_{ij}^{(k)} - \bar{Y}_i^{(k)})^2 / (N_k M_k - N_k)$ - істинна дисперсія між елементами усередині одиниць k-ї страти.

(S_{1k}^2 - аналог $\frac{S_c^2}{M}$ у кожній страті, а S_{2k}^2 - S_W^2, див. 5.3).

$s_{1k}^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\bar{y}_i^{(k)} - \bar{y}^{(k)})^2$ - незміщена оцінка S_{1k}^2,

$s_{2k}^2 = \frac{1}{n_k (m_k - 1)} \sum_{i=1}^{n_k} \sum_{j=1}^{m_k} \left(y_{ij}^{(k)} - \bar{y}_i^{(k)} \right)^2$ - незміщена оцінка S_{2k}^2.

$\tau^{(k)} = \sum_{i=1}^{N_k} \sum_{j=1}^{M_k} Y_{ij}^{(k)} = M_k \sum_{i=1}^{N_k} \bar{Y}_i^{(k)} = N_k M_k \mu^{(k)}$ - сумарне значення елементів у k-й страті;

$\tau = \sum_{k=1}^{L} N_k M_k \mu^{(k)}$ - сумарне значення елементів усієї сукупності;

$\mu = \frac{\tau}{\sum_{j=1}^{L} N_j M_j}$ - середнє значення елементів сукупності.

Тоді $\mu = \sum_{k=1}^{L} W_k \mu^{(k)}$, де

$W_k = N_k M_k / \left(\sum_{i=1}^{L} N_i M_i \right)$ - відносна вага страти в термінах елементів.

Незміщеною оцінкою μ буде статистика

$$\bar{\bar{y}}_{st} = \sum_{k=1}^{L} W_k \bar{y}^{(k)},$$

з дисперсією [9, 13] $D\bar{\bar{y}}_{st} = \sum_{k=1}^{L} W_k^2 \left(\frac{1-f_{1k}}{n_k} S_{1k}^2 + \frac{1-f_{2k}}{n_k m_k} S_{2k}^2 \right)$,

де $f_{1k} = \frac{n_k}{N_k}$, $f_{2k} = \frac{m_k}{M_k}$.

Незміщеною оцінкою цієї дисперсії буде статистика [13]

$$\upsilon\left(\bar{\bar{y}}_{st}\right) = \sum_{k=1}^{L} W_k^2 \left(\frac{1-f_{1k}}{n_k} s_{1k}^2 + \frac{1-f_{2k}}{n_k m_k} s_{2k}^2 \right).$$

Незміщеною оцінкою τ буде статистика

$$\hat{\tau}_{st} = \sum_{k=1}^{L} N_k M_k \bar{y}^{(k)} = \bar{\bar{y}}_{st} \cdot \left(\sum_{k=1}^{L} N_k M_k \right).$$

$$D\hat{\tau}_{st} = \left(\sum_{k=1}^{L} N_k M_k \right)^2 \cdot D\bar{\bar{y}}_{st} \text{ і } \upsilon\left(\hat{\tau}_{st} \right) = \left(\sum_{k=1}^{L} N_k M_k \right)^2 \cdot \upsilon\left(\bar{\bar{y}}_{st}\right).$$

Контрольні запитання, вправи

Контрольні запитання й завдання

1. *Дайте визначення одноступінчастого кластерного відбору.*

2. *У яких випадках одноступінчастий кластерний відбір дає точнішу оцінку ніж простий випадковий відбір?*

3. *Які оцінки для середнього та сумарного значень сукупности можна використовувати у випадку, коли кластери неоднакового розміру?*

4. *Дайте визначення двоступінчастого кластерного відбору.*

5. *Які обсяги вибірок повинні бути на кожному етапі двоступінчастого кластерного відбору, для того, щоб похидка оціки була найменшою при фіксованих витратах?*

6. *Які статистики використовуються для оцінювання середнього та сумарного значень сукупності при двоступінчастому стратифікованому відборі?*

Вправи для самостійної роботи

1. Бухгалтерська фірма зацікавлена в оцінці рівня похибок (помилок) у ревізіях, які вона проводить. Сукупність містить 800 позивів. Фірма провела аудит 85 позивів (використовуючи просту випадкову вибірку). Кожен з цих позивів містить 215 запитів, які повністю перевірені на помилки. Один позив мав 4 помилки серед 215 запитів, ще один позов -–3 помилки, 4 позики мали – по дві помилки, 22 позива мали по одній помилці, і решта 57 позивів не мали помилок.

а) Оцінити середнє число помилок в усіх 800 позивах. Знайти оцінку дисперсії;

б) Оцінити сумарне число помилок у сукупності.

2. Дано 10 кластерів і загальне число елементів в усіх кластерах – 100. Проводиться одноступінчатий кластерний відбір з $n = 3:\ \ \tau_1 = 4,\ M_1 = 5;\ \tau_2 = 12,\ M_2 = 20;\ \tau_3 = 7,\ M_3 = 10$.

а) Знайти незміщену оцінку сумарного значення сукупності;

б) Оцінити дисперсію цієї оцінки;

в) Знайти оцінку по відношенню сумарного значення сукупності і оцінку дисперсії.

3. Контролер перевіряє кількість недоліків у телевізорах у 580 контейнерах. Кожен контролер містить 24 телевізора. Двоступінчата вибірка – 12 контейнерів і 3 телевізора, у кожному з вибраних контейнерів дала наступний результат:

Ящики	1	2	3	4	5	6	7	8	9	10	11	12
1	1	4	0	1	2	0	2	3	4	3	3	0
2	0	2	1	2	0	1	2	0	3	1	0	0
3	2	1	2	0	3	0	1	2	0	0	2	0

а) Оцінити середнє число недоліків у телевізорі, та дисперсію цієї оцінки;

б) Оцінити загальне число недоліків у всій партії телевізорів;

в) Нехай вартість обстеження одного контейнера $c_1 = 5$ грн., а одного телевізора - $c_2 = 10$ грн. На обстеження виділили 2000 грн. Які оптимальні обсяги вибірки? $(C = c_1 n + c_2 nm)$;

г) Нехай $V = 10$. Які повинні бути оптимальні обсяги вибірок, щоб витрати були мінімальними?

4. Фірма бажає знати, чи підтримують службовці нову інвестиціну політику. Фірма має 87 окремих філій, що розташовані по всій Україні. Оскільки необхідно терміново прийняти рішення, керівництво фірми використало одноступінчатий кластерний відбір. Використовуючи просту випадкову вибірку, було вибрано 15 філій і проведено суцільне опитування. Результати наведені у таблиці. Оцініть сумарне число службовців, які підтримують нову політику, оцініть пропорцію службовців, які підтримують нову політику. Обчислити оцінки дисперсій.

Філія	Число службовців	Число підтримуючих нову політику
1	51	42
2	62	53
3	49	40
4	73	45
5	101	63

6	48	31
7	65	38
8	49	30
9	73	54
10	61	45
11	58	51
12	52	29
13	65	46
14	49	37
15	55	42

5. Мерія міста зацікавлена в оцінці середніх витрат сімей на комунальні послуги за місяць. Було використано одноступінчатий кластерний відбір. У місті 60 мікрорайонів (кластерів). Випадкова вибірка 20 мікрорайонів і суцільне опитування у кожному вибраному мікрорайоні дало наступні результати

Вибраний район	Число сімей	Сумарні витрати на ком. послуги (грн.)
1	55	2210
2	60	2390
3	63	2430
4	58	2380
5	71	2760
6	78	3110
7	69	2780
8	58	2370
9	52	1990
10	71	2810
11	73	2930
12	64	2470
13	69	2830
14	58	2370
15	63	2390
16	75	2870
17	78	3210
18	51	2430
19	67	2730
20	70	2880

Оцінити середні витрати сім’ї на комунальні послуги та стандартну похибку.

6. Райдержадміністрація хоче оцінити число дерев із визначеним захворюванням у лісі. Ліс поділений на 10 великих зон (кластерів). Кожен кластер містить участки однакової площі. Чотири бригади лісників здатні провести обстеження, яке повинно бути зроблено за один день. За допомогою

двоступінчатого випадкового відбору вибрано 4 кластери і 6 участків в кожному з них. Кожна бригада може обстежити 6 участків. Дані про хворі дерева наведені у таблиці.

Кластер	Число участків	Число обстежених участків	Число хворих дерев на участках
1	12	6	15, 14, 21, 13, 9, 10
2	15	6	4, 6, 10, 9, 8, 5
3	14	6	10, 11, 14, 10, 9, 15
4	21	6	8, 3, 4, 1, 2, 5

Оцінити сумарне (загальне) число хворих дерев в районі.

7. Сукупність містить $N = 10$ одиниць, кожна з яких складається з $M_i = 6$ елементів. За допомогою двоступінчатого відбору вибрано 2 одиниці, по 3 елементи з кожної одиниці. Отримані значення характеристики елементів: 7, 5, 3 для першої одиниці; 4, 2, 3 – для другої одиниці.

а) Оцінити середнє значення сукупності μ.

б) Оцінити дисперсію $D\bar{y}$.

РОЗДІЛ 5

ЗАСТОСУВАННЯ МЕТОДИКИ КОМПЛЕКСНОГО ДОСЛІДЖЕННЯ ПРИ ПРОВЕДЕННІ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ВИБІРКОВИХ ДОСЛІДЖЕНЬ

Найбільш великі дослідження включають різні методики, обговорені у попередніх розділах. Вибіркове дослідження може бути стратифікованим з різними ступінями кластеризації з використанням оцінок по відношенню та по регресії із залученням інших змінних.

Формули для оцінки стандартних похибок чи дисперсій можуть стати занадто громіздкими, особливо якщо використовуються багатоступінчасті кластерні оцінки без повернення. Для того, щоб спростити ці питання, використовуються вибіркові ваги та ефекти проектування в комплексних дослідженнях. Ці питання будуть обговорені у даному розділі. Також зупинимося на оцінюванні дисперсії при комплексному дослідженні.

5.1. Збирання компонентів проекту

У попередніх розділах були описані основні компоненти комплексного дослідження: простий випадковий відбір, стратифікація і кластерний відбір. Спробуємо тепер об’єднати їх в один вибірковий проект. Наведемо зараз основні поняття у модульній формі.

1. Кластерний відбір з поверненням. Здобувається вибірка n кластерів з поверненням; i-й кластер вибирається з ймовірністю ψ_i. Оцінюємо сумарне

значення елементів у i-му класі за допомогою незміщеної оцінки $\hat{\tau}_i$. Далі покладемо n значень $u_i = \dfrac{\hat{\tau}_i}{\psi_i}$ як спостереження.

Оцінюємо сумарне значення сукупності через середнє $\bar{u}$ і оцінюємо дисперсію цієї оцінки через s_u^2 / n.

2. Кластерний відбір без повернення. Здобувається вибірка n кластерів без повернення; i-й кластер попадає у вибірку з ймовірністю π_i. Оцінюємо сумарне значення i-го кластера за допомогою незміщеної оцінки $\hat{\tau}_i$ і обчислюємо оцінку її дисперсії $\upsilon(\hat{\tau}_i)$. Використовуючи оцінку Горвіца-Томпсона $\hat{\tau} = \sum_{i=1}^{n} \dfrac{\hat{\tau}_i}{\pi_i}$.

3. Стратифікація. Нехай $\hat{\tau}_1,...,\hat{\tau}_L$- незміщені оцінки сумарних значень у стратах $\tau_1,...,\tau_L$ і $\upsilon(\hat{\tau}_1),...,\upsilon(\hat{\tau}_L)$- незміщені оцінки дисперсії.

Тоді $\hat{\tau} = \sum_{i=1}^{L} \hat{\tau}_i$, $\upsilon(\hat{\tau}) = \sum_{i=1}^{L} \upsilon(\hat{\tau}_i)$.

Зауважимо, що кластерний і стратифікований відбори можуть бути організовані один в одному.

Оцінки можна записати, використовуючи вибіркові ваги. Причому, вибіркова вага одиниці – обернена величина до ймовірності того, що ця одиниця потрапить до вибірки.

Наприклад, для стратифікованого відбору

$\hat{\tau}_{st} = \sum_{k=1}^{L} \sum_{j=1}^{n_k} w_{kj} y_{kj}$, де $w_{kj} = \dfrac{N_k}{n_k}$.

Оскільки $\sum_{k=1}^{L} \sum_{j=1}^{n_k} w_{kj} = N$, то $\bar{y}_{st} = \dfrac{\sum_{k=1}^{L} \sum_{j=1}^{n_k} w_{kj} y_{kj}}{\sum_{k=1}^{L} \sum_{j=1}^{n_k} w_{kj}}$.

Для двоступінчастого рівноімовірнісного відбору

$$\hat{\tau} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} y_{ij}, \quad \hat{\mu} = \frac{\hat{\tau}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}}.$$

Вся інформація, необхідна для побудови оцінок міститься у вибіркових вагах. Але вибіркові ваги не дають інформації як знайти дисперсію чи стандартну похибку оцінки. Дисперсії оцінок залежать від ймовірностей того, що будь-яка пара одиниць потрапила до вибірки. Методи оцінки дисперсій складних вибіркових планів будуть описані в наступних параграфах.

5.2. Ефект проекту

В роботі [15] запропоновано виміряти ефективність вибіркового плану через відношення дисперсії оцінки, отриманої при використанні комплексного вибіркового плану з n обстеженими одиницями до дисперсії оцінки, отриманої за допомогою простого випадкового відбору n одиниць. Леслі Кіш у своїй книзі [L.Kish, 21], назвав це відношення ефектом проекту (design effect) вибіркового плану.

Наприклад, для оцінки середнього значення сукупності по вибірці обсягу n одиниць

$$deff\left(plan,\bar{y}\right)=\frac{D\hat{\mu}_p}{\left(1-\frac{n}{N}\right)\frac{S^2}{n}},$$

де $D\hat{\mu}_p$ - дисперсія оцінки, отриманої за допомогою вибіркового плану.

Ефект проекту забезпечує міру точності, отриману чи втрачену за рахунок використання комплексного проекту замість простої випадкової вибірки.

Наприклад, для пропорційного стратифікованого відбору

$$deff\left(\bar{y}_{st},\bar{y}\right)=\frac{D\bar{y}_{st}}{D\bar{y}}\approx\frac{\sum_{k=1}^{L}W_kS_k^2}{S^2}\approx\frac{\sum_{k=1}^{L}W_kS_k^2}{\sum_{k=1}^{L}W_kS_k^2+\sum_{k=1}^{L}W_k\left(\mu_k-\mu\right)^2}.$$

Якщо всі μ_k різні, то $deff\left(\bar{y}_{st},\bar{y}\right)<1$. Тобто, стратифікація дає більшу точність на одиницю спостережень, ніж проста випадкова вибірка.

Для одноступінчастого кластерного відбору при великих N

$$deff\left(\hat{\mu}_{claster},\bar{y}\right)=deff\left(\hat{\tau}_{claster},\hat{\tau}_{SRS}\right)\approx1+\rho_c\left(M-1\right).$$

Якщо коефіцієнт кореляції $\rho_c > 0$, то ефект проекту більший одиниці. Тобто кластерний відбір в цьому випадку дає меншу точність на одиницю спостережень, ніж проста випадкова вибірка.

Якщо ж, $\rho_c < 0$, то, можливо, що $deff\left(\hat{\mu}_{claster}, \bar{y}\right) < 1$ і кластерний відбір дасть більшу точність на одиницю спостережень ніж проста випадкова вибірка.

5.3. Оцінювання дисперсії в комплексних дослідженнях

У попередніх розділах були отримані формули дисперсій для різних вибіркових планів. Деякі формули досить прості, як, наприклад, для простої випадкової вибірки. А, скажімо, обчислення дисперсії чи її оцінки при двоступінчастому кластерному відборі без повернення є досить громіздким.

У цьому параграфі будуть описані декілька методів оцінювання дисперсій при комплексних вибіркових дослідженнях.

1. Метод лінеаризації

У попередніх розділах оцінки сумарного значення чи середнього значення були лінійною комбінацією оцінених сумарних значень чи середніх у стратах чи кластерах.

Нехай

$$\hat{\tau} = \sum_{i=1}^{L} a_i \hat{\tau}_i \,. \qquad (5.1)$$

Тоді

$$D\hat{\tau} = D\left(\sum_{i=1}^{L} a_i \hat{\tau}_i \right) = \sum_{i=1}^{L} a_i^2 D\hat{\tau}_i + 2\sum_{i=1}^{L} \sum_{j=i+1}^{L} a_i a_j \operatorname{cov}\left(\hat{\tau}_i, \hat{\tau}_j\right). \qquad (5.2)$$

Аналогічна формула справедлива і для дисперсії середнього значення.

Якщо ж ми використовуємо оцінки по відношенню, то вони вже не будуть лінійними комбінаціями типу (5.1).

Нехай $h\left(\hat{\tau}_1, ..., \hat{\tau}_L\right)$ - оцінка τ і функція h - нелінійна, тоді, використовуючи формулу Тейлора для функції h

$$h(t_1,...,t_L) \approx a_0 + \sum_{i=1}^{L} a_i t_i \ .$$

Отже, $Dh(\hat{\tau}_1,...,\hat{\tau}_L)$ можна апроксимувати $D\left(\sum_{i=1}^{L} a_i \hat{\tau}_i\right)$ і використати формулу (5.2).

Як визначити константи a_i? За формулою Тейлора

$$h(\hat{\tau}_1,\hat{\tau}_2,...,\hat{\tau}_L) \approx h(\tau_1,\tau_2,...,\tau_L) + \sum_{j=1}^{L} a_j (\hat{\tau}_j - \tau_j),$$

де $a_j = \left.\dfrac{\partial h(t_1,...,t_L)}{\partial t_j}\right|_{t_i=\tau_i,\ i=\overline{1,L}}$. Покладемо $q_i = \sum_{j=1}^{L} a_j y_{ij}$.

Далі шукаємо дисперсію оцінки $\hat{\tau}_q = \sum_i w_i q_i$, де w_i - відповідні ваги. Це і є метод лінеаризації.

Приклад 5.1. У параграфі 2.7 середньоквадратичне відхилення $E(\hat{R} - R)^2$ апроксимували виразом

$$\frac{1-f}{n\mu_x^2}\sum_{i=1}^{N}(Y_i - RX_i)^2 / (N-1) = \frac{1-f}{n\mu_x^2}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y\right) =$$
$$= \frac{(1-f)N^2}{n\tau_x^2}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y\right) = \frac{1}{\tau_x^2}\left(D(\hat{\tau}_y) + R^2 D(\hat{\tau}_x) - 2R\,\mathrm{cov}(\hat{\tau}_x,\hat{\tau}_y)\right) \qquad (5.3)$$

Ми використали, що

$$D(\hat{\tau}_y) = \frac{N^2(1-f)}{n} S_y^2,\ D(\hat{\tau}_x) = \frac{N^2(1-f)}{n}\cdot S_x^2,$$

$$\mathrm{cov}(\hat{\tau}_x,\hat{\tau}_y) = \frac{N^2(1-f)}{n}\cdot \rho S_x S_y \ \text{ і } \ f = \frac{n}{N}.$$

Використовуючи метод лінеаризації, отримаємо цю формулу.

$$\hat{R} = \frac{\hat{\tau}_y}{\hat{\tau}_x},\ R = \frac{\tau_y}{\tau_x} = h(\tau_x,\tau_y).$$

Нехай $h(t_1,t_2) = \dfrac{t_2}{t_1}$. $\dfrac{\partial h(t_1,t_2)}{\partial t_1} = -\dfrac{t_2}{t_1^2}$, $\dfrac{\partial h(t_1,t_2)}{\partial t_2} = \dfrac{1}{t_1}$. $a_1 = -\dfrac{\tau_y}{\tau_x^2}$, $a_2 = \dfrac{1}{\tau_x}$.

Тоді

$$\hat{R} - R \approx -\frac{\tau_y}{\tau_x^2}\left(\hat{\tau}_x - \tau_x\right) + \frac{1}{\tau_x}\left(\hat{\tau}_y - \tau_y\right)$$

та

$$E\left(\hat{R} - R\right)^2 \approx E\left(-\frac{\tau_y}{\tau_x^2}\left(\hat{\tau}_x - \tau_x\right) + \frac{1}{\tau_x}\left(\hat{\tau}_y - \tau_y\right)\right)^2 =$$

$$= \frac{\tau_y^2}{\tau_x^4} D\hat{\tau}_x + \frac{1}{\tau_x^2} D\hat{\tau}_y - 2\frac{\tau_y}{\tau_x^3} \cdot \operatorname{cov}\left(\hat{\tau}_x, \hat{\tau}_y\right) =$$

$$= \frac{1}{\tau_x^2}\left(R^2 D\hat{\tau}_x + D\hat{\tau}_y - 2R\operatorname{cov}\left(\hat{\tau}_x, \hat{\tau}_y\right)\right)$$

і отримали формулу (5.3).

Далі для оцінки цього виразу треба обчислювати оцінки $\hat{R}, \upsilon\left(\hat{\tau}_x\right), \upsilon\left(\hat{\tau}_y\right)$, оцінку коваріації і, можливо, оцінку $\hat{\tau}_x$ (якщо невідомо τ_x). Альтернативно можна обчислити $q_i = \frac{1}{\tau_x}\left(y_i - \hat{R}x_i\right)$ і шукати $\upsilon\left(\hat{\tau}_q\right)$.

2. Метод випадкових груп

Нехай оцінюється сумарне значення сукупності по вибірці обсягу n якимось методом. Тоді випадковим чином розбиваємо вибірку на s груп, кожна розміром n/s і s разів обчислюємо оцінку (в кожній групі). Маємо послідовність оцінок $\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_s$.

Тоді $\bar{\hat{\mu}} = \frac{1}{s}\sum_{i=1}^{s} \hat{\mu}_i$ і $\upsilon_1\left(\hat{\mu}\right) = \frac{1}{s(s-1)}\sum_{i=1}^{s}\left(\hat{\mu}_i - \bar{\hat{\mu}}\right)^2$.

Останній вираз можна і використовувати як оцінку $\upsilon\left(\hat{\mu}\right)$.

Альтернативною може бути і оцінка

$$\upsilon_2\left(\hat{\mu}\right) = \frac{1}{s(s-1)}\sum_{i=1}^{s}\left(\hat{\mu}_i - \hat{\mu}\right)^2.$$

3. Джекнайф-метод

Джекнайф-метод (Jackknife method, JK-метод) розширює метод випадкових груп, дозволяючи групам перекриватися. Вперше цей метод увів Куеноуілл [Quenouill, 28, 29] як метод зменшення зсуву оцінки. У книзі [34] досить детально

описаний цей метод. Розглянемо JK-метод на прикладі відкидання однієї одиниці (хоча цей метод дозволяє відкидати довільноможливу кількість одиниць).

Нехай $\hat{\tau}_{(j)}$- оцінка τ(аналогічно підрахована, як і оцінка $\hat{\tau}$), але без j-го спостереження.

Наприклад, для простої випадкової вибірки,

$$\hat{\tau}_{(j)} = N \cdot \bar{y}_{(j)} = N \cdot \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} y_i .$$

Далі $\upsilon_{JK}(\hat{\tau}) = \frac{n-1}{n} \sum_{j=1}^{n} \left(\hat{\tau}_{(j)} - \hat{\tau}\right)^2$.

У нашому прикладі

$$\hat{\tau}_{(j)} = \frac{N}{n-1}\left(\sum_{i=1}^{n} y_i - y_j\right) = \frac{N}{n-1}\left(n\bar{y} - y_j\right) = N\left(\bar{y} - \frac{1}{n-1}\left(y_j - \bar{y}\right)\right).$$

Тоді $\sum_{j=1}^{n}\left(\hat{\tau}_{(j)} - \hat{\tau}\right)^2 = \sum_{j=1}^{n}\left(\hat{\tau}_{(j)} - N\bar{y}\right)^2 = \frac{N^2}{(n-1)^2}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2 = \frac{N^2}{(n-1)} s_y^2$

і $\upsilon_{JK}(\hat{\tau}) = \frac{N^2}{n} s_y^2$, що є оцінкою дисперсії $\hat{\tau}$ при простому випадковому відборі з поверненням. При стратифікованому відборі з поверненням JK-метод застосовується окремо в кожній страті, тоді

$\upsilon_{JK}(\hat{\tau}_{st}) = \sum_{k=1}^{L} \frac{n_k - 1}{n_k} \sum_{j=1}^{n_k}\left(\hat{\tau}_{(kj)} - \hat{\tau}_{st}\right)^2$, де $\hat{\tau}_{(kj)}$- оцінка сумарного значення у k-й страті і без j-го вибраного елемента.

При кластерному відборі JK-метод можна використовувати на першому етапі відбору, тобто j - це номер вибраної одиниці (кластера).

В основному цей метод використовується при рівноімовірнісному відборі з поверненням.

4. Будстреп-метод

Як і JK-метод, будстреп-метод (Boodstrap method, BS-метод) був розвинутий для різних напрямків статистики, інших ніж техніка вибіркових досліджень. В книзі

[34] узагальнені теоретичні результати BS-методу для комплексних вибіркових досліджень. Розглянемо застосування цього методу, на прикладі простої випадкової вибірки без повернення. Отже, маємо вибірку обсягу n.

Робимо далі N/n копій цієї вибірки і, таким чином, "розтягуємо" вибірку до розміру сукупності N ("псевдосукупність" чи "псевдопопуляція"). Далі вибираємо s простих випадкових вибірок без повернення з цієї "псевдосукупності" обсягу n і користуємось оцінками методу випадкових груп. Якщо n/N мале, то BS-розподіли з поверненням чи без повернення подібні. Опишемо BS-метод для стратифікованого відбору з поверненням [31].

1. Здобуваємо просту випадкову вибірку з поверненням обсягу $n_k - 1$ в k-й страті.

2. Для кожного $r(r = 1,2,...,s)$ обчислюємо нову вагу i-ї одиниці k-ї страти

$$w_{ki}(r) = w_{ki} \cdot \frac{n_k}{n_k - 1} m_{ki}(r),$$

де $m_{ki}(r)$ - частота (кількість) появи i–го елемента у вибірці. Обчислюємо оцінку $\hat{\tau}_{st}^{*}(r)$, використовуючи ваги $w_{ki}(r)$.

3. Повторюємо кроки 1 та 2 s разів.

4. Обчислюємо $\upsilon_{RS}(\hat{\tau}) = \frac{1}{s-1} \sum_{r=1}^{s} \left(\hat{\tau}_{st}^{*}(r) - \hat{\tau}\right)^2$.

Детальний опис застосування BS-методу для оцінки дисперсій при вибіркових дослідженнях можна знайти також в роботах [17, 18, 25, 36].

Метод лінеаризації широко використовується для оцінки дисперсій в комплексних вибіркових дослідженнях. Головний недолік цього методу в тому, що потрібно знати всі частинні похідні функції h.

Метод випадкових груп досить простий і привабливий для використання. Його можна використовувати для будь-якого вибіркового плану. Головний недолік у тому, що для стійких оцінок потрібно, щоб число груп s було великим. А при обмеженому обсягу вибірки n це не завжди можна зробити.

JK-метод та BS-метод також широко застосовуються, але вони вимагають більших обмежень, ніж попередні методи.

Розроблено декілька комп'ютерних пакетів програм, які допомагають при аналізі данних комплексних вибіркових досліджень. Найбільш відомий пакет SUDAAN. Усі описані методи оцінювання дисперсії реалізовані в цьому пакеті [14, 22].

5.4. Використання вибірки для вивчення динаміки соціально-економічних явищ

При періодичному повторенні обстежень з метою вивчення динаміки соціально-економічних явищ (бюджети домашніх господарств, фінансова діяльність фермерських господарств, ділова активність промислових підприємств, банків, торгових підприємств, динаміка цін, безробіття і зайнятість, стан навколишнього середовища, стан здоров'я населення і таке інше) можуть бути використані чотири вида вибірки:

1) незалежна (змінна);
2) постійна (фіксована);
3) ротаційна (часткове заміщення);
4) підвибірка.

Незалежна (змінна) вибірка – це вибіркове обстеження, яке повторюється через визначені проміжки часу. При цьому кожен раз нова вибірка здобувається незалежно від попередніх, склад і число одиниць відбору є змінною величиною.

Постійна (фіксована) вибірка - повторне вибіркове обстеження, що проводиться за тією ж самою вибіркою. Одиницею відбору виступає фіксована на даний період часу група об'єктів вибірки. Проте у зв'язку з тим, що протягом деякого проміжку часу можуть відбутися зміни в даній зафіксованій групі об'єктів (створення нових одиниць відбору, ліквідація існуючих і таке інше) має сенс щорічна адаптація фіксованої вибірки до змін, що відбуваються.

Ротаційна вибірка (часткове зміщення) припускає при проведенні повторного вибіркового дослідження здійснення заміни лише частини початкової

вибірки. При багатократному повторенні можна скласти план заміщення. Наприклад, кожен раз замінювати четверту частину вибірки; при цьому кожна відібрана одиниця залишається у вибірці в чотирьох наступних обстеженнях.

Підвибірка означає повторне обстеження підвибірки з початкової вибірки.

Кожна з розглянутих вибірок має свої переваги і недоліки.

Нехай період повтору обстеження – рік. Незалежна вибірка за кожен рік достатньо точно відображає склад і структуру початкової сукупності, але може дати неспівставні результати. Це обумовлено двома факторами : 1) вибірки кожного року проводяться незалежно одна від одної, що ускладнює розрахунок показників динаміки; 2) кожен рік виникають похибки оцінювання, причому різними по величині, що приводить до того, що рівень явища може виявитися випадково значно високим одного року (рік t), а у наступному році ($t+1$) він може бути недооціненим. В результаті динаміка досліджуємого явища викривляється. При використанні фіксованої вибірки виходять з того, що похибка оцінок в році t приблизно співпадає з похибкою оцінок в рік ($t+1$). В результаті динаміка і розвиток досліджуємого процесу оцінюється більш точно, але вимагається щорічна адаптація постійної вибірки. Ротаційна вибірка поєднує в собі елементи змінної і постійної вибірок, проте вона характеризує структуру генеральної сукупності менш точно, ніж незалежна, і дає менш співставні результати, ніж фіксована.

Метою багатократних вибірок може бути:

а) оцінка зміни сумарного значення τ і середнього значення μ від одного моменту часу до іншого;

б) оцінка середнього значення τ або μ за весь період;

в) оцінка значення μ в кожен із моментів обстеження.

У першому випадку необхідно стежити за розвитком явища, тому краще всього у різні моменти часу мати одну і ту ж вибірку, тобто використовувати постійну вибірку, у другому випадку рекомендується незалежно здобуті вибірки, у третьому – віддають перевагу комбінації двох схем або ротаційній вибірці.

При здобутті вибірок через деякі визначені проміжки часу виникають ряд проблем:

- повторне обстеження одних і тих же одиниць може виявитись практично незручним, бо визиває небажання надавати необхідні відомості (збільшується навантаження респондентів);

- на опитуваних може впливати інформація, отримана в ході опитування, що з часом може зменшити репрезентативність поданих відомостей;

- повторні обстеження можуть привести до зміни одиниць спостереження у порівнянні із іншою частиною сукупності;

- в основі вибірки з часом відбуваються зміни (з'являються нові одиниці сукупності, зникають старі; наприклад, створюються нові підприємства, ліквідуються збиткові);

- виникає проблема періодичної адаптації постійної вибірки, розробка ефективного механізму заміщення вибірки.

Правила актуалізації вибірки

При формуванні і адаптації (актуалізації) фіксованої вибірки її стратифікація, послідовність груп (страт) і вибіркових долей не повинні змінюватися у порівнянні з попереднім роком. Об'єкти, які знаходилися зовні сукупності в минулому році (нові підприємства, нові домогосподарства і таке інше) формують окрему страту. В кожну страту (групу) бажано відбирати об'єкти, які знаходилися у вибірці раніше. Але якщо розмір вибірки вищий чи нижчий необхідного, тоді деяке число об'єктів може бути виключене з вибірки або, навпаки, додатково включено до неї. При цьому необхідно враховувати кількість років, протягом яких кожен об'єкт був включеним або не був включеним в обстеження. Усім одиницям у сукупності надають номери обстеження. Величина номера показує , скільки років обстежується дана одиниця. При здійсненні відбору об'єкти спостереження розташовуються не у випадковому порядку, а послідовно, у відповідності з величиною номера обстеження. Об'єкти з однаковими номерами розміщуються довільно. В результаті навантаження розподіляється відносно рівномірно між усіма респондентами.

У зарубіжній статистичній практиці [6, 13] розроблено ряд правил актуалізації основи фіксованої вибірки:

1. Актуалізується число і основні ознаки одиниць спостереження. Наприклад, якщо обстежуються підприємства деякої галузі, то враховуються зміни їх розміру і основного виду діяльності.

2. За період актуалізації основних ознак одиниці спостереження зазвичай вибирається один рік. Це пов'язано з тим, що внесення щомісячних змін в основу вибірки дуже ускладнює процес використання річних даних, проведення спостережень і аналіз. Також дуже важко вказати місяць виникнення змін, що стосуються розміру і основного виду діяльності тих чи інших резидентів. Подібні зміни відбуваються поступово, реальні зміни можуть відбутися значно раніше вказаного терміну, інколи часовий розрив може складати навіть два роки.

3. Зміна числа обстежуваних одиниць відображається не щорічно, а по ходу утворення нових і ліквідації існуючих об'єктів дослідження (щомісячно). В основі вибірки з моменту першого запису в ній враховуються усі фактори створення і ліквідації одиниць спостереження. Виникає певний елемент "відставання" реєстрації від реальної ситуації. Проте всі фактори створення і ліквідації враховуються в показниках динаміки. При цьому згладжуються різкі підйоми і падіння і більш чітко відстежуються довготривалі тенденції розвитку.

За допомогою фіксованої вибірки розв'язуються задачі:

1) оцінка річних абсолютних рівнів досліджуємих показників;

2)оцінка місячних абсолютних рівнів досліджуємих показників;

3)сумісна оцінка місячних і річних абсолютних рівнів досліджуємих показників;

4) забезпечення співставлення місячних і річних показників;

5) забезпечення співставлення порівнювальних показників;

6) можливість побудови довготривалих динамічних рядів економічних показників.

Задачі і принципи формування ротаційної і фіксованої вибірок співпадають. В результаті змін, які відбуваються з часом у сукупності, проводиться адаптація вибірки. Особливість ротаційної вибірки заключається в тому, що кількість об'єктів, якими повинні бути доповнені підгрупи чи які повинні бути виключені з них,

визначаються заздалегідь. Розроблюється спеціальний план заміщення вибіркових одиниць. Об‘єктами, які виключаються із обстеження, виступають ті, що більш триваліший строк (до 5 років) були включені у спостереження. Навантаження на респондентів, що приймають участь у обстеженні протягом тривалого часу, скорочується.

Контрольні запитання, вправи

Контрольні запитання й завдання

1. *У чому полягає суть комплексних досліджень?*

2. *Дайте визначення ефекту ппроету.*

3. *Які основні методи оцінювання дисперсії в комплексних дослідженнях?*

4. *Які особливості застосування методу лінеаризації?*

5. *У яких випадках застосовується метод випадкових груп, джекнайф метод і будстреп метод?*

6. *Дайте визначення незалежній вибірці, постійній вибірці та ротаційній вибірці.*

7. *Назвіть правила актуалізації вибірки.*

Вправи для самостійної роботи

1. Коефіцієнт кореляції сукупності є

$$\rho = \frac{\sum_{i=1}^{N}(X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(X_i - \mu_x)^2 \sum_{i=1}^{N}(Y_i - \mu_y)^2}} = \frac{\tau_{xy} - \frac{\tau_x \cdot \tau_y}{N}}{\sqrt{\tau_{x^2} - \frac{(\tau_x)^2}{N}}\sqrt{\tau_{y^2} - \frac{(\tau_y)^2}{N}}}.$$

$R = h(t_1, t_2, t_3, t_4, t_5)$, де $t_1 = \tau_x, t_2 = \tau_y,\ t_3 = \tau_{x^2},\ t_4 = \tau_{xy},\ t_5 = \tau_{y^2}$.

Нехай $\hat{\rho} = h\left(\hat{\tau}_x, \hat{\tau}_y, \hat{\tau}_{x^2}, \hat{\tau}_{xy}, \hat{\tau}_{y^2}\right)$. Всі оцінки $\hat{\tau}_x, \hat{\tau}_y, \hat{\tau}_{x^2}, \hat{\tau}_{xy}, \hat{\tau}_{y^2}$ незміщені.

а) Використовуючи метод лінеаризації, знайти апроксимацію дисперсії для дисперсії оцінки $\hat{\rho}$.

б) Яка буде апроксимація дисперсії для простої випадкової вибірки без повернення обсягу n?

2. Підрахувати ефект проекту для систематичної вибірки. Зробити аналіз.

Рекомендована література

1. Анісімов В.В., Черняк О.І. Математична статистика.- К.: МП “Леся”, 1995. – 104 с.

2. Бокун Н.Ч., Чернышева Т.М. Методы выборочных обследований: Учебно-справочное пособие. – Мн., 1997. – 416 с.

3. Вибіркове спостереження : Термінологічний словник / О.О.Васєчко, О.І.Черняк, Є.М.Жуйкова та інші. – К.: Державний комітет статистики України, 2004. – 140 с.

4. Гладун О.М., Саріогло В.Г. Застосування процедур імпутації відсутніх даних вибіркового обстеження умов життя домогосподарств методами “середнього значення”, “пропорцій”, “hot-deck” . Методичні матеріали. - К.: Державний комітет статистики України, 2001. – 23 с.

5. Джессен Р. Методы статистических обследований. – М.: Финансы и статистика, 1985. – 478 с.

6. Йейтс Ф. Выборочный метод в переписях и обследованиях – М.:Статистика, 1965.- 435 с.

7. Литл Р. Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками. - М.: Финансы и статистика, 1990. –336 с.

8. Рыжова Т. Возможности применения выборочого метода при оценке основних средств как элементов національного багатства // Вопросы экономики.- 1993.- N 5.- С.116-119.

9. Черняк О.І. Техніка вибіркових досліджень. – К.: МІВВЦ, 2001. - 248 с.

10. Черняк О.І., Обушна О.М., Ставицький А.В. Теорія ймовірностей та математична статистика. Збірник задач: Навч.посіб. 2-ге вид. – К.: Знання, КОО, 2002. - 199 с.

11. Черняк О.І. Застосування теорії оптимального розміщення при проведенні маркетингових досліджень // Вісник Донецького університету. Серія В. Економіка і право. – 2003.- N 1. – С. 55-58.

12.Черняк О.І. Оптимальна частка відбору серед респондентів, які не відповіли при вибірковому обстеженні // Теоретичні та прикладні питання економіки.-2004. - N 4. – С.120-128.

13.Cochran W.G. Sampling technique, 3 ed. –New York: John Wiley & Sons, 1977. – 411 p.

14.Cohen S.B. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data // American Statistician. – 1997. – N 51. - P. 285-292.

15.Cornfield J. Modern methods in the sampling of human populations // American Journal of Public Health. – 1951. – N 41. - P. 654-661.

16.Deming W.E. Some theory of sampling. – New York: John Wiley & Sons, 1950. – 603 p.

17.Efron B. Bootstrap methods: Another look at the jackknife // Annals of Statistics. – 1979. - Vol.7. - P. 1-26.

18.Efron B. The jackknife, the bootstrap, and other resampling plans. – Philadelphia: SIAM, 1982. – 278 p.

19.Hansen M.M., Hurwitz W.N. The problem of nonresponse in sample surveys // Journal of American Statistical Association. – 1946. - Vol.41. - P. 517-529.

20.Hansen M.M., Hurwitz W.N., Madow W.G. Sample survey methods and theory (in two volumes). – New York: John Wiley & Sons, 1953. – 970 p. (I –638 p., II – 332 p.).

21.Kish L. Survey sampling. 2 ed. – New York: John Wiley & Sons, 1976. – 642 p.

22.Lehtonen R., Pahkinen E.J. Practical methods for design and analysis of complex survey. – New York: John Wiley & Sons, 1996. – 320 p.

23.Lessler J., Kalsbeek W. Nonsampling errors in surveys. – New York: John Wiley & Sons, 1992. – 412 p.

24.Levy P.S., Lemeshow S. Sampling of populations.: Methods and applications. – New York: John Wiley & Sons, 1991. – 386 p.

25.Lohr S.L. Sampling: Design and analysis. – New York: Duxbury Press., 1999. – 484 p.

26.Neyman J. On the two different aspcts of the representative method: The method of stratified sampling and the method of purposive selection // Journal of the Royal Statistical Society. – 1934. - Ser. A. - Vol. 97. - P. 558-625.

27.Neyman J. Contribution to the theory of sampling human population // Journal of the American Statistical Association. – 1938. - Vol. 33. - P. 101-116.

28.Quenouille M.H. Problems in plane sampling // Annals of Mathematical Statistics. – 1949. - Vol. 20. - P. 355-375.

29.Quenouille M.H. Notes on bias in estimation // Biometrika. – 1956. – N 43. - P. 353-360.

30.Rao J.N.K. On double sampling for stratification and analytical surveys // Biometrika. – 1973. - Vol. 60. – N 1. - P. 125-133.

31.Rao J.N.K., Wu C.F.J. Resampling inference with complex survey data // Journal of the American Statistical Association. – 1988. - Vol. 83. - P. 231-241.

32.Rao J.N.K. Some current trends in sample survey theory and methods // The Indian Journal of Statistics. – 1999. - Vol. 61. - Series B. - N.1. - P. 1-57.

33.Särndal C.-E., Swensson B., Wretman J. Model assisted survey sampling. – New York: Springer – Verlag, 1992. - 694 p.

34.Shao J., Tu D. The jackknife and bootstrap. – New York: Springer-Verlag, 1995. – 420 p.

35.Thompson M.E. Theory of sample surveys. – London: Chapman & Hall, 1997. – 300 p.

36.Tryfos P. Sampling methods for applied research. – New York: John Wiley & Sons, 1996. - 440 p.

СЛОВНИК

Вибірка (вибіркова сукупність) - група чи частина одиниць спостереження , відібраних за певними правилами із генеральної сукупності.

Вибірка багатоступінчаста - вибірка, процес формування якої проходить кілька послідовних етапів чи ступенів; спочатку з генеральної сукупності відбираються групи великих одиниць (кластери), потім з великих – середні, потім – дрібні й серед останніх здійснюється вибір окремих одиниць, що підлягають спостереженню. На заключному етапі одиниця відбору збігається з одиницею спостереження.

Вибірка двоступінчаста (двоетапна) - вибіркова сукупність, сформована в два етапи: на першому етапі у випадковому порядку вибираються одиниці (кластери), що підлягають обстеженню, на другому етапі з кожної відібраної одиниці у випадковому порядку відбирається певна кількість елементів (підодиниць), що підлягають безпосередньому спостереженню тобто:

1. Вибирається n одиниць із N одиниць (проста випадкова вибірка);

2. Вибирається m_i елементів (підодиниць) із вибраної i-ї одиниці (кластера) (проста випадкова вибірка обсягу m_i із сукупності в M_i одиниць).

Вибірка кластерна - це імовірнісна вибірка, у якій кожна вибіркова одиниця є групою або кластером елементів, які не перетинаються та разом охоплюють усю сукупність, що обстежується. У відібраних кластерах можуть обстежуватися або всі елементи кластера, або їх частина (багатоступінчаста вибірка). Якщо у відібраних кластерах обстеженню підлягають усі без винятку одиниці, то вибірка називається одноступінчастою кластерною , а основа, із якої вона формується, первинною основою вибірки. Якщо ж у відібраних на першому ступені кластерах проводиться подальший підвибір одиниць (тобто спостерігається тільки частина елементів кластера), то вибірка називається двоступінчастою кластерною, а її основа повторною і т.п.

Вибірка комплексна - вибірка заснована на комбінації різних видів вибірки. Наприклад, стратифікований відбір з різними ступенями кластеризації.

Вибірка одноступінчаста - вибірка, при якій випадковим чином вибираються одиниці (кластери), і всі елементи (підодиниці) кластеру підлягають обстеженню.

Вибірка проста випадкова - вибірка, що складається із випадково відібраних одиниць із генеральної сукупності, у якій кожна одиниця має однакову ймовірність бути включеною до вибірки або кожна можлива комбінація однакового числа одиниць із генеральної сукупності має однакову ймовірність утворити вибірку. Заснована на випадковому відборі одиниць спостереження із основи вибірки без усякого розчленування її на частини або групи. Для організації вибірки використовується метод жеребкування, таблиця випадкових чисел, комп'ютерний датчик "випадкових чисел".

Вибірка репрезентативна (представницька) - вибірка, яка адекватно відображає систему ознак обстежуваної сукупності. Мірою репрезентативності вибірки є похибки репрезентативності вибірки.**Вибірка систематична** - вибірка, заснована на систематичному (через певний інтервал) відборі з генеральної сукупності, одиниці якої розташовуються у певному порядку (за алфавітом, географічним принципом, у порядку зростання чи зменшення значень якої-небудь ознаки). Підмножина складається із систематично відібраних одиниць із несистематизованої або систематизованої основи вибірки. Проміжок, через який одиниці попадають у вибірку, залежить від прийнятої пропорції відбору. Вона встановлюється діленням чисельності сукупності на обсяг вибірки. Перша одиниця вибірки вибирається випадково.

Вибірка стратифікована - вибірка, що включає ряд вибірок, узятих з відповідних страт генеральної сукупності. Вона складається із визначеної кількості випадково відібраних одиниць частин досліджуваної сукупності, які являють собою однорідні страти (групи) відносно досліджуваних ознак. Страти утворюються таким чином, щоб одиниці усередині кожної з них були якнайбільш схожими між собою (мали невелику або помірну варіацію усередині групи). Відбір одиниць у кожній із утворених страт проводиться або у випадковому порядку, або шляхом

систематичного відбору (за умови достатньої кількості одиниць у страті). Основне призначення стратифікованої вибірки полягає у тому, щоб за рахунок стратифікації сукупності одержати більш високу точність результатів вибірки у порівнянні із простим випадковим відбором при тому ж обсязі вибіркової сукупності, або ту ж точність при меншому обсязі вибірки. Стратифікована вибірка має широке застосування при вивченні неоднорідних сукупностей, коли ні випадкова, ні систематична вибірка не можуть застосовуватись внаслідок дуже високої варіації системи досліджуваних ознак, асиметричності розподілу їхніх значень.

Вибірковий метод - система правил відбору одиниць і способів характеристики сукупності досліджуваних одиниць, що вивчаються. Дозволяє розповсюдити висновки, що отримані на основі вивчення частини сукупності (вибірки) на всю генеральну сукупність .

Вибіркові одиниці - частини основи вибірки, що являються елементами відбору, по яким робляться висновки про властивості всієї популяції.

Відбір багатоступінчастий - відбір, під час якого процес формування сукупності вибіркової проходить декілька послідовних етапів, або ступенів. На перших ступенях відбираються значні групи (кластери), із них відповідно до плану вибіркового спостереження знову провадиться відбір кластерів, але вже більш дрібних, і так доти, поки на останньому етапі не будуть відібрані одиниці, які підлягають обстеженню. На останньому етапі одиниця відбору збігається із одиницею спостереження. Багатоступінчастий відбір може бути двоступінчастим, триступінчастим і т.д. Основні переваги багатоступінчастого відбору полягають: по-перше, у порівняльній нескладності утворення основи вибірки; по-друге, у можливості використання як існуючого природного розподілу досліджуваного матеріалу на кілька частин різних рангів, так і штучно утворених одиниць; по-третє, у можливості складати основу вибірки на наступних ступенях лише для тих одиниць першого ступеня, що потрапили до вибірки. У той же час багатоступінчастий відбір дає загалом менш точні результати, ніж одноступінчастий відбір тієї ж кількості одиниць останнього ступеня.

Відбір без повернення - відбір (вибірка без повернення), при якому один раз відібрана одиниця виключається із подальшого відбору і тому не може бути відібрана повторно.

Відбір випадковий - відбір, під час якого ще до його здійснення кожна одиниця генеральної сукупності володіє визначеною, заздалегідь заданою (у найпростішому випадку – однаковою) імовірністю бути включеною до вибірки. Випадковий відбір може бути з поверненням і без повернення. Для організації випадкового відбору використовуються або таблиці випадкових чисел, або спосіб жеребкування.

Відбір двоступінчастий (двоетапний) - відбір, при якому вибірка здійснюється в два етапи: на першому випадковим чином здійснюється вибірка одиниць (кластерів), що підлягають обстеженню , на другому – з кожної відібраної одиниці здійснюється випадкова вибірка елементів.

Відбір із імовірностями, пропорційними оцінці розміру - спосіб відбору одиниць при формуванні багатоступінчастої вибірки, коли відомий лише приблизний розмір первинних одиниць-кластерів (кількість елементів, що містяться в них).

Відбір імовірнісний - відбір, заснований на об'єктивних правилах випадкового відбору, під час якого кожний елемент сукупності має відому ненульову ймовірність бути відібраним до вибірки. При імовірнісному відборі завжди існує можливість точно зазначити як множину різних вибірок $S_1, S_2, ..., S_n$, які можуть бути отримані за даним методом відбору із конкретної сукупності, так і відповідні одиниці відбору, що належать сформованим вибіркам.

Відбір кластерний - відбір при вибірковому спостереженні, що полягає в тому, що відбираються не окремі одиниці, а цілі групи (кластери) чи серії. Основою вибірки є сукупність кластерів (серій або груп елементів), що не перетинаються, та разом вичерпують генеральну сукупність.

Відбір комплексний - відбір, під час якого включення одиниць до вибірки здійснюється шляхом сполучення різних його способів (наприклад випадковий і

кластерний, систематичний і кластерний, багатоступінчастий, кластерний і систематичний і т.д.).

Відбір нерівноімовірнісний - відбір, під час якого кожна елементарна одиниця (або група одиниць) має неоднакову ненульову імовірність потрапити до вибірки, тобто i-та одиниця вибирається з імовірністю π_i, $i = 1,2,...,N$.

Відбір одноступінчастий - відбір, при якому випадковим чином вибираються одиниці (кластери), і всі елементи (підодиниці) підлягають обстеженню.

Відбір з поверненням - відбір, при якому відібрані у вибірку один раз одиниці не виключаються з генеральної сукупності і подальшого процесу відбору і можуть бути відібраними повторно.

Відбір первинних одиниць з рівними імовірностями - спосіб відбору вихідних одиниць при формуванні багатоступінчастої вибірки (первинні одиниці-кластери відбираються з однаковою імовірністю, незалежно від їхнього розміру)

Відбір подвійний (двофазний) - спосіб, при якому за деякими ознаками, що цікавлять дослідника, аналізується сукупність на підставі вивчення всіх її одиниць, а за іншими ознаками – на підставі частини одиниць сукупності вибіркової, яка представляє підвибірку із одиниць первісної вибірки.

Відбір подвійний (двофазний) для оцінки по відношенню - із сукупності генеральної випадковим чином відбирається вибірка із n' елементів (перша фаза) і в цій вибірці проводиться спостереження тільки ознаки x; потім із n' елементів у другій фазі випадковим чином відбирається n елементів, по яких спостерігаються ознаки і X, і Y, величина Y оцінюється по відношенню.

Відбір подвійний (двофазний) для оцінки по регресії - із генеральної сукупності, що містить N елементів, випадковим чином відбирається допоміжна вибірка з n' елементів (перша фаза); для цих n' елементів спостерігається ознака X, яка відносно легко чи дешево піддається виміру і корелює з досліджуваною ознакою Y; з допоміжної вибірки робиться основна підвибірка, що складається з n елементів (друга фаза), у рамках якої проводяться основні виміри Y, величина Y оцінюється по регресії.

Відбір подвійний (двофазний) для стратифікації - відносно велика вибірка з n' елементів випадковим чином здобувається із генеральної сукупності, що містить N елементів (перша фаза); для цих елементів виміряється ознака X, і на підставі отриманої інформації вибірка поділяється на L страт; потім з кожної страти випадковим чином здобувається вибірка, для елементів якої вимірюється ознака Y (друга фаза).

Відбір, пропорційний розміру страт (пропорційне розміщення) - тип стратифікованого відбору, при якому розподіл обсягу вибірки за різними стратами здійснюється пропорційно питомій вазі кожної групи в загальній сукупності ($f_k = n_k / N_k = n/N, k = 1,2..., L,$ де L- кількість страт, N- кількість одиниць в генеральній сукупності, N_k - кількість одиниць у k-ій страті, n- загальний обсяг вибірки, n_k- обсяг вибірки із k-ї страти). Частка відбору однакова для всіх страт.

Відбір простий випадковий - відбір одиниць обстеження з сукупності генеральної випадковим способом без попереднього розподілення генеральної сукупності на будь-які групи, при цьому будь-яка одиниця вибирається з однаковою імовірністю. При **простому випадковому відборі без повернення** ймовірність вибору будь-якої індивідуальної вибірки S із n одиниць із сукупності з N одиниць буде $\frac{1}{C_N^n} = \frac{n!(N-n)!}{N!}$, де $n! = 1 \cdot 2 \cdot ... \cdot (n-1) \cdot n$, а ймовірність того, що i-та одиниця потрапить до вибірки, дорівнює $\pi_i = \frac{n}{N}$, тобто вона однакова для всіх одиниць. При **простому випадковому відборі з поверненням** кожна одиниця вибирається з ймовірністю $\frac{1}{N}$.

Відбір систематичний - відбір при вибірковому спостереженні, суть якого полягає в тому, що складається список одиниць генеральної сукупності і у залежності від числа одиниць (серій), що відбираються, встановлюється крок відбору, тобто через який інтервал варто брати для спостереження одиниці (серії). Наприклад, до вибірки включаються кожна 5, 10, 15, 20 і т.д. одиниці (крок відбору дорівнює 5). Початок відбору визначається або за таблицею випадкових чисел із

номерів, які відповідають першому інтервалу (наприклад, 1, 2, 3, 4, 5), або способом жеребкування.

Відбір стратифікований - це спосіб формування вибірки з урахуванням структури генеральної сукупності. Цей відбір передбачає попередню структуризацію генеральної сукупності і незалежний відбір одиниць у кожній складовій, яку називають стратою. При стратифікованому відборі генеральна сукупність ділиться на страти (типові групи одиниць) за будь-якою ознакою. Страти не мають спільних одиниць і разом вичерпують усю генеральну сукупність. А потім з кожної із страт проводиться систематичний відбір чи випадковий відбір.

Відношення – це величина відношення двох сумарних значень різних характеристик (ознак) генеральної сукупності $R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i}$.

Генеральна сукупність (популяція) - сукупність одиниць або елементів, про які необхідно зробити певні висновки.

Досліджувана сукупність - сукупність одиниць спостереження, що складає об'єкт дослідження і підлягає статистичному вивченню.

Ефект проекту - відношення дисперсії оцінки, отриманої при використанні комплексного вибіркового плану з n-обстеженими одиницями до дисперсії оцінки, отриманої за допомогою простого випадкового відбору n одиниць. Ефект проекту $(deff)$ забезпечує міру точності, отриману $(deff < 1)$ чи втрачену $(deff > 1)$ за рахунок використання комплексного відбору замість простої випадкової вибірки.

Імпутація - це процедура заповнення відсутніх значень по окремих ознаках або по групах ознак, які вимірюються за програмою обстеження .

Метод будстреп (BS-метод) – метод оцінювання дисперсії в комплексних дослідженнях, пов'язаний із „розтягуванням" вибірки до розміру сукупності і використанням методу випадкових груп.

Метод випадкових груп - метод оцінювання дисперсії в комплексних дослідженнях, пов'язаний із розбиттям вибірки на випадкові групи і оцінювання дисперсії в кожній групі.

Метод джекнайф (JK-метод)- метод оцінювання дисперсії в комплексних дослідженнях, пов’язаний із відкиданням одиниць вибірки і усередненням оцінок дисперсії по багатьом вибіркам.

Метод лінеаризації - метод оцінювання дисперсії в комплексних дослідженнях, пов’язаний із лінійною апроксимацією оцінки дисперсії.

Об′єкт спостереження (статистичного) - сукупність явищ (у статистиці - сукупність статистична), що підлягають статистичному спостереженню. Це може бути сукупність підприємств, жителів країни і т.д. Для успішного проведення спостереження об'єкт спостереження має бути чітко визначеним. Для цього на основі аналізу досліджуваного явища потрібно виділити і вказати ознаки та риси, що відрізняють його від інших подібних з ним об'єктів, визначити межі переходу від одного явища до іншого. Цьому допомагає точне визначення одиниці спостереження.

Обсяг вибірки - кількість об’єктів (одиниць) у вибірковій сукупності (n). Це можуть бути або окремі одиниці, або групи їх (кластери, серії), відібрані із генеральної сукупності. Від обсягу вибірки. залежить репрезентативність результатів вибіркового обстеження. Чим більше обсяг вибірки, тим менше помилка репрезентативності, точніші дані вибіркового спостереження.

Обсяг сукупності - чисельність одиниць, що складають статистичну сукупність.

Одиниця відбору - одне з основних понять вибіркового методу, складовий елемент (одиниця) генеральної сукупності чи їх група (кластер, серія, природні групи одиниць спостереження (будинки, квартири)) послідовним відбором яких формується вибіркова сукупність. Одиниця відбору за своїми розмірами не повинна бути меншою одиниці спостереження, тому що це приводить до систематичної помилки репрезентативності.

Одиниця кластерна вибіркова - група елементів, що у процесі формування вибірки розглядається як одна одиниця. У простому випадку елементи, що складають кластер (серію) або входять у вибірку як група, або не входять до неї взагалі.

Одиниця спостереження (ОС) - первинний елемент об’єкту статистичного спостереження, що характеризується сукупністю ознак.

Одиниця сукупності - кожний окремо взятий елемент даної генеральної сукупності , що має певні ознаки. Це складовий елемент об'єкта спостереження, ознаки якого повинні бути вимірені, оцінені під час обстеження (в іноземній літературі використовується термін “елементарна одиниця”), наприклад, при демографічних обстеженнях одиницею спостереження може бути людина, але може бути і домогосподарство, у залежності від того, яка мета ставиться перед обстеженням. При вивченні фінансово-економічної діяльності підприємств одиницею спостереження є підприємство. Для кожного статистичного спостереження одиниця повинна бути чітко визначена, названі її ознаки.

Ознака – значення або характеристика одиниці спостереження, яка підлягає реєстрації в процесі статистичного спостереження. Наприклад, якщо одиницею спостереження є промислове підприємство, то його ознаками можуть бути: середньо-облікова кількість штатних працівників облікового складу, вартість основних виробничих фондів, а також форма власності, належність до того чи іншого виду економічної діяльності тощо.

Основа вибірки (обстежувана сукупність) - сукупність одиниць, які підлягають вивченню, і система її визначення . Або це опис виду одиниць, із яких складається сукупність, і виклад правил включення або не включення будь-якої окремої одиниці до складу даної сукупності. Основа вибірки нерозривно пов'язана з одиницями відбору, вона може бути визначена як перелік, список одиниць відбору. Якщо одиниця відбору збігається з одиницею спостереження, то основа вибірки адекватна генеральній сукупності. Якщо одиниця відбору об'єднує багато одиниць спостереження (серія, кластер), то основа вибірки відрізняється від генеральної сукупності. При багатоступінчастому відборі на кожному ступені своя основа вибірки. Будь-яка основа вибірки повинна бути точною, повною і вільною від подвійного рахування, відповідати цілям дослідження як на момент її побудови, так і на момент її використання. Старіння основи вибірки за час між моментами її побудови і використання може призводити до помилок у результатах вибірки.

Оцінка - показник, побудований за визначеним правилом за результатами вибірки; приймається як замінник справжнього (невідомого) показника чи значення, яке характеризує генеральну сукупність.

Оцінка незміщена – це оцінка, у якої її очікуване значення по всім можливим вибіркам згідно даного типу відбору (математичне сподівання) дорівнює істинному значенню сукупності.

Оцінка роздільна (по відношенню, по регресії) - одержання роздільних оцінок параметрів генеральної сукупності для кожної страти (групи) за даними стратифікованої вибірки і наступне їхнє підсумовування.

Оцінка комплексна (по відношенню, регресійна) - регресійна оцінка, використання якої пов'язане із застосуванням додаткової інформації від ознаки, кореляційно пов'язаної з оцінюваною.

Оцінка сумісна (по відношенню, по регресії) - одержання оцінки параметрів генеральної сукупності за даними стратифікованої вибірки, виходячи з єдиного сумісного відношення (оцінка по відношенню) чи єдиного сукупного коефіцієнта регресії (оцінка по регресії).

Оцінка стратифікована - оцінка, побудована на основі стратифікованого відбору: $\overline{\mu}_{st} = \sum_{i=1}^{L} w_i \overline{\mu}_i$, w_i - вага i-ї страти, $\overline{\mu}_i$ - оцінка характеристики в i-й страті.

План вибірки - організаційно-логічна модель структури вибіркової сукупності.

Похибка вибіркова (репрезентативності) – різниця між істинним значенням статистичного показника та його значенням, розрахованим у вибірці. Найчастіше використовують такі показники похибки вибірки: "стандартна похибка", "коефіцієнт варіації".

Похибка спостереження - систематична похибка, що виникає при проведенні складних обстежень. Джерела похибок спостереження: навмисний витяг "представницької" вибірки; залежність процесу відбору від якої-небудь ознаки, пов'язаної з досліджуваними властивостями одиниці; недотримання принципу випадковості відбору; не отримання відповідей; похибки реєстрації. За своїм складом похибка спостереження складається з декількох неоднорідних компонент: постійного зміщення, змінної складової зміщення, випадкової складової похибки.

Програма спостереження - перелік ознак одиниці спостереження, які підлягають реєстрації у процесі проведення спостереження.

Пропорція – частка одиниць в сукупності, які мають деяку властивість (атрибут).

Реєстр статистичний - список складових частин (одиниць) об'єкта статистичного спостереження. У статистичному реєстрі відзначаються зміни (поява нових, вибуття старих) його складових частин (наприклад, реєстр населення, реєстр будівництв). Статистичний реєстр є основою реєстрової форми спостереження. Реєстр статистичних одиниць (у статистиці України) – це інформаційна система, яка забезпечує визначення сукупності одиниць статистичного спостереження, накопичення та збереження даних про одиниці спостереження, пошук інформації про кожний об’єкт, за яким ведеться спостереження. У статистичному реєстрі містяться всі облікові одиниці, одиниці спостереження або аналітичні одиниці із відповідними ознаками, які необхідні для проведення статистичних спостережень.

Розміщення одиниць нейманівське оптимальне - розміщення, яке враховує ступінь варіації ознаки у різних стратах генеральної сукупності. Розрахунок обсягу вибірки з кожної страти проводиться за формулою: $n_k = n \cdot \frac{N_k S_k}{\sum_{j=1}^{L} N_j S_j} = n \cdot \frac{W_k S_k}{\sum_{j=1}^{L} W_j S_j}$, $k = 1,...,L$, де L - кількість страт, N_k - кількість одиниць у k-ій страті, S_k^2 - дисперсія k-ї страти, n_k - обсяг вибірки із k-ї страти, n - загальний обсяг вибірки. $n = n_1 + n_2 + ... + n_L$, W_k - вага k-ї страти.

Розміщення одиниць пропорційне - розміщення, при якому кількість одиниць, що відбираються у вибірку, пропорційна питомій вазі даної групи за числом одиниць генеральної сукупності, тобто число спостережень кожної страти визначається за формулою: $n_k = n\frac{N_k}{N}$, де n_k - число спостережень з k-тої страти, n - загальний обсяг вибірки, N_k - розмір k-ї страти, N - розмір генеральної сукупності.

Розповсюдження (екстраполяція) - розрахунок даних генеральної сукупності за вибірковими даними.

ДОДАТКИ

Статистичні таблиці

Таблиця А.1

Статистична таблиця. Випадкові числа

74970	06996	11136	26528	23607	97462
74077	63454	45058	20708	42772	61311
13557	72942	59693	42635	69187	17870
66824	77092	51315	11910	91362	85877
36135	62333	37762	06766	52006	48746
06176	37697	40726	66014	78540	03503
17371	29089	26149	86755	36502	45455
21223	60124	07325	61085	61663	93814
31842	75317	58670	07821	75722	75152
20516	27594	21126	21262	14847	85513

99277	64548	70107	01059	34794	89863
01991	83000	27894	43577	82087	71504
54377	90482	39785	75722	20978	72511
20121	24555	25752	35312	85403	46189
11571	25668	34005	60874	72564	27470
93725	16472	21779	22432	71132	58118
65299	19900	21083	77915	20234	57314
36671	66533	86361	01327	80226	67405
49870	72912	20126	71728	86130	22113
50647	27134	56117	08650	91732	56189
17834	90311	00470	25024	20604	55526
27421	59467	69163	36665	26139	59445
26586	93561	52994	91112	74191	53986
51769	19891	46105	60143	63230	43817
41635	22882	85301	06875	58116	90778
04382	75863	37867	86246	58449	47432
48736	95362	21908	86094	43262	82826
49226	85080	33783	98388	62526	04014
20854	80874	15061	24566	72654	83590
50093	79411	58243	12538	16000	81354
32746	91894	87531	03933	08670	35011
45655	67247	49062	80256	21828	70217
96268	69668	23518	85192	81640	19832
43792	70776	17047	10233	44527	40725
66726	38354	88229	52784	48167	43464
00305	60732	03985	83552	83744	33572
47203	23522	41528	72453	88184	97289
94417	00980	76255	09103	55746	57149
28492	27329	28987	08292	22457	27594
15068	78906	13085	52751	42272	10144
86628	62686	03694	38080	35208	10638
70099	52095	34944	74139	92323	24202
59642	03751	88891	73720	90197	48857
21373	68891	89516	31394	29618	13531
62249	55787	68112	51338	09111	84084
15068	28465	20985	64222	79260	22767
35078	08613	30709	07408	99171	30553
19643	91937	12828	53404	07541	10589
75025	72481	37200	27222	92688	11164
71553	58597	83573	12991	32797	24758

Таблиця А.2

Значення t_α для розподілу Стьюдента залежно від імовірності $P\{|t_k| < t_\alpha\} = 1-\alpha$ і числа степенів свободи k. Щільність розподілу дорівнює:

$$P_{t_k}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \cdot \left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

$1-\alpha$ k	0,90	0,95	0,99
1	6,31	12,71	63,7
2	2,92	4,30	9,92
3	2,35	3,18	5,84

4	2,13	2,77	4,60
5	2,02	2,57	4,03
6	1,943	2,45	3,71
7	1,895	2,36	3,50
8	1,860	2,31	3,36
9	1,833	2,26	3,25
10	1,812	2,23	3,17
11	1,796	2,20	3,11
12	1,782	2,18	3,06
13	1,771	2,16	3,01
14	1,761	2,14	2,98
15	1,753	2,13	2,95
16	1,746	2,12	2,92
17	1,740	2,11	2,90
18	1,734	2,10	2,88
19	1,729	2,09	2,86
20	1,725	2,09	2,84
25	1,708	2,06	2,79
30	1,697	2,04	2,46
80	1,659	1,991	2,640
100	1,651	1,984	2,627
∞	1,645	1,960	2,576

Таблиця А.3

Значення c_α.

$1-\alpha$	0,8	0,9	0,95	0,99	0,999
c_α	1,28	1,64	1,96	2,57	3,29

Розділ 4. Оцінювання по відношенню та по регресії

Франція не мала перепису населення у 1802 році і Лаплас захотів оцінити кількість населення у країні [19]. Він одержав вибірку із 30 общин по всій країні. Ці общини мали 2037615 жителів у вересні 1802 року. За три попередні роки до вересня 1802 р. в цих общинах було зареєстровано 215599 народжень дітей. Лаплас визначив щорічну кількість народжень у цих общинах: 215599/3=71866,33. Ділячи 2037615 на 71866,33 Лаплас оцінює, що кожного року було зареєстровано одне народження на 28,352845 осіб. Міркуючи, що будь-які общини з великим населенням також ймовірно будуть мати великі ряди зареєстрованих народжень, Лаплас припустив, що співвідношення населення до щорічних народжень у його вибірці повинно бути подібним і для всієї Франції.

Таким чином, він запропонував оцінити загальну кількість населення Франції за допомогою множення загальної кількості щорічних народжень у всій Франції на 28,352845. Лаплас не був зацікавлений загальною кількістю народжень, але використав цю інформацію для оцінки загальної кількості населення Франції.

Часто допоміжна інформація може бути використана для того, щоб покращити точність оцінок. У цьому розділі розглядаються оцінки по відношенню та по регресії, які використовують змінні, що корелюють із головною змінною для вдосконалення оцінок середнього та сумарного значень сукупності.

4.1. Оцінювання по відношенню

Нехай, як і у параграфі 2.7, i-та одиниця характеризується двома змінними X_i та Y_i. Причому допоміжна змінна X_i корельована з головною змінною Y_i. Усі значення X_i відомі. Через x_i та y_i позначимо вибіркові значення i–ї одиниці, що потрапила у вибірку, $i = 1, 2, ..., n$. Скористаємось позначеннями параграфа 2.7:

$$R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{\tau_y}{\tau_x} = \frac{\tau_y / N}{\tau_x / N} = \frac{\mu_y}{\mu_x}$$ - відношення,

$\hat{R} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \dfrac{\bar{y}}{\bar{x}}$ - оцінка відношення.

Тоді оцінка по відношенню середнього значення сукупності Y буде наступною

$$\hat{\mu}_R = \hat{R} \cdot \mu_x = \frac{\bar{y}}{\bar{x}} \cdot \mu_x . \qquad (4.1)$$

А оцінка сумарного значення сукупності Y

$$\hat{\tau}_R = N \cdot \hat{\mu}_x = \hat{R} \cdot \tau_x = \frac{\bar{y}}{\bar{x}} \cdot \tau_x .$$

Оскільки оцінки по відношенню $\hat{\mu}_R$ та $\hat{\tau}_R$ зміщені, то для порівняння з іншими незміщеними оцінками необхіно підрахувати середньоквадратичні відхилення, тобто

$$MSE\left(\hat{\mu}_R\right) = E\left(\hat{\mu}_R - \mu\right)^2; \quad MSE\left(\hat{\tau}_R\right) = E\left(\hat{\tau}_R - \tau\right)^2 .$$

Відомо, що для зміщених оцінок справедлива рівність

$$MSE\left(\hat{\mu}_R\right) = D\hat{\mu}_R + \left(E\hat{\mu}_R - \mu\right)^2$$

$$MSE\left(\hat{\tau}_R\right) = D\hat{\tau}_R + \left(E\hat{\tau}_R - \tau\right)^2 .$$

Другий доданок у цих формулах – квадрат зсуву оцінки [1]. Причому цей доданок для оцінок по відношенню менший по зрівнянню з дисперсією, тому наближені формули для середнього квадратичного відхилення такі ж самі, як і для дисперсії.

Справедливі наступні наближені формули для середньоквадратичного відхилення чи дисперсії оцінок по відношенню середнього та сумарного значень сукупності при великому обсягу вибірки n [18].

$$D\hat{\mu}_R \approx \frac{1-f}{n}\sigma_R^2, \quad \text{де } \sigma_R^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - RX_i\right)^2 , \qquad (4.2)$$

$$D\hat{\tau}_R \approx N^2\,\frac{1-f}{n}\sigma_R^2 . \qquad (4.3)$$

Відповідно оцінками даних дисперсій будуть статистики

$$\upsilon(\hat{\mu}_R) = \frac{1-f}{n}s_R^2, \text{ де } s_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \hat{R}x_i\right)^2, \ \upsilon(\hat{\tau}_R) = N^2\,\frac{1-f}{n}s_R^2 .$$

$1-\alpha$ -надійні інтервали для середнього та сумарного значень сукупності відповідно знаходяться за формулами:

$$\mu_y \in \left(\hat{\mu}_R - t_\alpha\sqrt{\upsilon(\hat{\mu}_R)};\ \hat{\mu}_R + t_\alpha\sqrt{\upsilon(\hat{\mu}_R)}\right)$$
$$\tau_y \in \left(\hat{\tau}_R - t_\alpha\sqrt{\upsilon(\hat{\tau}_R)};\ \hat{\tau}_R + t_\alpha\sqrt{\upsilon\left(\hat{\tau}_R\right)}\right) .$$

де t_α – табличне значення розподілу Стьюдента з $n-1$ степеню свободи. При $n > 50$ t_α можна замінити на c_α (див.табл.2, табл.3 додатка).

Нехай ρ - коефіцієнт кореляції X та Y, тобто

$$\rho = \frac{\sum_{i=1}^{N}(Y_i - \mu_y)(X_i - \mu_x)}{(N-1)S_y S_x} \qquad (4.4)$$

де $S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(X_i - \mu_x)^2$, $S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(Y_i - \mu_y)^2$ - дисперсії величин X та Y.

Отримаємо еквівалентну формулу для $D\hat{\mu}_R$.

$$D\hat{\mu}_R \approx \frac{1-f}{n(N-1)}\sum_{i=1}^{N}\left[(Y_i - \mu_y) - R(X_i - \mu_x)\right]^2 =$$

$$= \frac{1-f}{n(N-1)}\left[\sum_{i=1}^{N}(Y_i - \mu_y)^2 + R^2\sum_{i=1}^{N}(X_i - \mu_x)^2 - 2R\sum_{i=1}^{N}(Y_i - \mu_y)(X_i - \mu_x)\right] =$$

$$= \frac{1-f}{n}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x\right).$$

Тут була використана рівність $\mu_y = R\mu_x$.

$$D\hat{\tau}_R \approx \frac{N^2(1-f)}{n}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x\right).$$

Порівняємо тепер дисперсії оцінки по відношенню і середнього простої випадкової вибірки

$$D\bar{y} - D\hat{\mu}_R = \frac{1-f}{n}\left(S_y^2 - S_y^2 - R^2 S_x^2 + 2R\rho S_y S_x\right) = \frac{1-f}{n}RS_x\left(2\rho S_y - RS_x\right).$$

Нехай $R > 0$, тоді $D(\bar{y}) \geq D(\hat{\mu}_R)$, якщо

$$\rho \geq \frac{RS_x}{2S_y} = \frac{1}{2}\left(\frac{S_x}{\mu_x}\right)\Big/\left(\frac{S_y}{\mu_y}\right) = \frac{c_x}{2c_y},$$

де c_x, c_y - коефіцієнти варіації X та Y. Якщо ж $c_x \approx c_y$, то $\rho \geq 0{,}5$. Таким чином , кореляція сукупностей X та Y повинна бути не слабою.

4.2. Оцінювання по відношенню при стратифікованому відборі: роздільні та сумісні оцінки

Нехай сукупність, що досліджується, розбивається на L страт (див.розділ 3): i-та одиниця k-ї страти має дві характеристики Y_{ki} та X_{ki}. Вважаємо, що величини x_{ki} всі відомі. Для оцінки середнього значення та сумарного значення сукупності по Y можна побудувати дві оцінки по відношенню: роздільну та сумісну (separate and combined) [18].

Роздільні оцінки по відношенню

У кожній страті обчислюємо оцінку відношення і результати сумуємо (seperate ratio).

Тобто $\widehat{\mu}_{R_S} = \sum_{k=1}^{L} W_k \cdot \frac{\bar{y}_k}{\bar{x}_k} \cdot \mu_{x_k}$, де $\bar{y}_k$, $\overline{x_k}$ - вибіркові середні у k-й страті, μ_{x_k} - істинне середнє в k-й страті,

$$\mu_{x_k} = \frac{1}{N_k} \sum_{i=1}^{N_k} X_{ki},$$

$$\widehat{\tau}_{R_s} = \sum_{k=1}^{L} \frac{\bar{y}_k}{\bar{x}_k} \cdot \tau_{x_k} = N \widehat{\mu}_{R_s}, \text{ де } \tau_{x_k} = N_k \cdot \mu_{x_k} = \sum_{i=1}^{N_k} X_{ki}.$$

Очевидно, що при великих обсягах вибірок n_k

$$D\left(\widehat{\mu}_{R_s}\right) = \sum_{k=1}^{L} W_k^2 \frac{(1-f_k)}{n_k} \left(S_{y_k}^2 + R_k^2 S_{x_k}^2 - 2R_k \rho_k S_{y_k} S_{x_k}\right), \text{ де } R_k = \frac{\mu_{y_k}}{\mu_{x_k}} = \frac{\tau_{y_k}}{\tau_{x_k}}$$ - істинне відношення у k-й страті. $S_{y_k}^2$, $S_{x_k}^2$ - дисперсії Y та X у k-й страті, ρ_k - коефіцієнт кореляції Y та X у k-й страті, $f_k = \frac{n_k}{N_k}$.

$$D\left(\widehat{\tau}_{R_s}\right) = \sum_{k=1}^{L} \frac{N_k^2 (1-f_k)}{n_k} \left(S_{y_k}^2 + R_k^2 S_{x_k}^2 - 2R_k \rho_k S_{y_k} S_{x_k}\right).$$

Сумісні оцінки по відношенню

У цьому випадку шукається лише одне відношення, сумісне для всіх страт (combined ratio). Нехай

$$\bar{y}_{st} = \sum_{k=1}^{L} W_k \bar{y}_k,\ \bar{x}_{st} = \sum_{k=1}^{L} W_k \bar{x}_k .$$

Тоді $\hat{\mu}_{R_c} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \cdot \mu_x$, де $\mu_x = \sum_{k=1}^{L} W_k \mu_{x_k}$.

$$\hat{\tau}_{R_c} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \cdot \tau_x = N \cdot \hat{\mu}_{R_c},\quad \tau_x = \sum_{k=1}^{L} \tau_{x_k} .$$

Якщо сумарний обсяг вибірки великий, то

$$D\hat{\mu}_{R_c} = \sum_{k=1}^{L} W_k^2 \frac{(1-f_k)}{n_k}\left(S_{y_k}^2 + R^2 S_{x_k}^2 - 2R\rho_k S_{y_k} S_{x_k}\right), \text{ де } R = \frac{\mu_y}{\mu_x},$$

$$\mu_y = \sum_{k=1}^{L} W_k \mu_{y_k},\ \mu_{y_k} = \frac{1}{N_k} \sum_{i=1}^{N_k} Y_{ki} .$$

$$D\hat{\tau}_{R_c} = \sum_{k=1}^{L} N_k^2 \frac{(1-f_k)}{n_k}\left(S_{y_k}^2 + R^2 S_{x_k}^2 - 2R\rho_k S_{y_k} S_{x_k}\right).$$

Порівняємо тепер дисперсії сумісної та роздільної оцінок:

$$D\hat{\mu}_{R_c} - D\hat{\mu}_{R_s} = \sum_{k=1}^{L} W_k^2 \frac{(1-f_k)}{n_k}\left[\left(R^2 - R_k^2\right)S_{x_k}^2 - 2(R-R_k)\rho_k S_{y_k} S_{x_k}\right] =$$

$$= \sum_{k=1}^{L} W_k^2 \frac{(1-f_k)}{n_k}\left[(R-R_k)^2 S_{x_k}^2 + 2(R_k - R)\left(\rho_k S_{y_k} S_{x_k} - R_k S_{x_k}^2\right)\right] .$$

В ситуаціях, коли застосовуються оцінки по відношенню, другий доданок у квадратних дужках за звичай малий. У більшості випадків, коли R_k дуже відрізняються між собою, роздільна оцінка буде більш точною.

4.3. Оцінювання по регресії

Оцінки по відношенню дають найкращий результат, коли величини Y та X лінійно залежні і лінія регресії проходить через початок координат. Припустимо тепер, що Y та X зв'язані довільною лінійною залежністю, або близькі до такої залежності: $Y = a + bX$. Далі використовуємо позначення параграфу 4.1. Вважаємо, що всі значення X_i відомі. Тоді оцінка по регресії середнього значення сукупності Y буде такою:

$$\hat{\mu}_l = \bar{y} + b\left(\mu_x - \bar{x}\right),$$

а оцінка сумарного значення сукупності Y - $\hat{\tau}_l = N\hat{\mu}_l$.

Зауважимо, що у випадках:

а) $b = 0$, то $\hat{\mu}_l = \bar{y}$ (середнє простої випадкової вибірки);

б) $b = \frac{\bar{y}}{\bar{x}}$, то $\hat{\mu}_l = \bar{y} + \frac{\bar{y}}{\bar{x}}\left(\mu_x - \bar{x}\right) = \frac{\bar{y}}{\bar{x}} \cdot \mu_x = \hat{\mu}_R$ (оцінка по відношенню).

Очевидно, що $\hat{\mu}_l$ та $\hat{\tau}_l$ – незміщені оцінки. Дійсно,

$$E\hat{\mu}_l = E\bar{y} + bE\left(\mu_x - \bar{x}\right) = \mu_y + b\left(\mu_x - \mu_x\right) = \mu_y .$$

Далі $E\hat{\tau}_l = NE\hat{\mu}_l = N \cdot \mu_y = \tau_y$.

Дисперсія оцінки $\hat{\mu}_l$ має вигляд [18]

$$D\hat{\mu}_l = \frac{(1-f)}{n(N-1)} \cdot \sum_{i=1}^{N}\left(\left(Y_i - \mu_y\right) - b\left(X_i - \mu_x\right)\right)^2 = \frac{1-f}{n}\left(S_y^2 - 2bS_{yx} + b^2 S_x^2\right),$$ де

$S_{yx} = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(Y_i - \mu_y\right)\left(X_i - \mu_x\right)$ – коваріація Y та X, $f = \frac{n}{N}$.

$$D\hat{\tau}_l = N^2 D\hat{\mu}_l = \frac{(1-f)N^2}{n}\left(S_y^2 - 2bS_{yx} + b^2 S_x^2\right).$$

Якою повинна бути константа b, щоб $D\hat{\mu}_l$ і $D\hat{\tau}_l$ були мінімальними?

Мінімум параболи $b^2 S_x^2 - 2bS_{yx} + S_y^2$ досягається у вершині

$$B = \frac{S_{yx}}{S_x^2} = \frac{\sum_{i=1}^{N}\left(Y_i - \mu_y\right)\left(X_i - \mu_x\right)\Big/(N-1)}{\sum_{i=1}^{N}\left(X_i - \mu_x\right)^2\Big/(N-1)}.$$

Ця величина називається коефіцієнтом лінійної регресії Y на X у скінченній сукупності.

Якщо ρ – коефіцієнт кореляції (див.(4.4)), то $B = \rho \cdot \frac{S_y}{S_x}$.

Надалі будемо використовувати лише оцінки $\hat{\mu}_l = \bar{y} + B\left(\mu_x - \bar{x}\right)$ і $\hat{\tau}_l = N\left(\bar{y} + B\left(\mu_x - \bar{x}\right)\right)$.

Тоді

$$D\hat{\mu} = \frac{1-f}{n}\left(S_y^2 - 2 \cdot \rho\frac{S_y}{S_x} \cdot S_{yx} + \rho^2 \frac{S_y^2}{S_x^2} \cdot S_x^2\right) = \frac{1-f}{n}\left(1-\rho^2\right)S_y^2, \text{ бо } \rho = \frac{S_{yx}}{S_x S_y}.$$

Оцінкою величини B може служити статистика (оцінка методу найменших квадратів, МНК-оцінка).

$$\hat{B} = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

А незміщеною оцінкою дисперсії $D\hat{\mu}_l$ –

$$\upsilon(\hat{\mu}_l) = \frac{(1-f)}{n(n-2)} \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{B}(x_i - \bar{x}) \right]^2 = \frac{(1-f)}{n(n-2)} \left\{ \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right\}.$$

$\upsilon(\hat{\tau}_l) = N^2 \upsilon(\hat{\mu}_l).$

$(1-\alpha)$-надійні інтервали для μ_y та τ_y відповідно знаходяться за формулами

$$\mu_y \in \left(\hat{\mu}_l - t_\alpha \sqrt{\upsilon(\hat{\mu}_l)};\ \hat{\mu}_l + t_\alpha \sqrt{\upsilon(\hat{\mu}_l)} \right);$$
$$\tau_y \in \left(N\hat{\mu}_l - Nt_\alpha \sqrt{\upsilon(\hat{\mu}_l)};\ N\hat{\mu}_l + Nt_\alpha \sqrt{\upsilon(\hat{\mu}_l)} \right),$$

де t_α – табличне значення розподілу Стьюдента з $n-2$ степенями свободи.

При $n > 50$ t_α можна замінити на c_α (див.табл.2 та 3 додатка).

Порівняємо тепер дисперсії оцінок по регресії, по відношенню та середнього простої випадкової вибірки. Припустимо, що n достатньо велике

$$V_1 = D\hat{\mu}_l = \frac{(1-f)}{n} S_y^2 (1-\rho^2);$$
$$V_2 = D\hat{\mu}_R = \frac{(1-f)}{n} \left(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y \right),$$
$$V_3 = \frac{(1-f)}{n} S_y^2 .$$

Очевидно, що $V_1 \geq V_3$ (рівність лише, коли $\rho = 0$)

$$V_2 - V_1 = \frac{(1-f)}{n} \left(R^2 S_x^2 - 2\rho S_x S_y + \rho^2 S_y^2 \right) = \frac{(1-f)}{n} (RS_x - \rho S_y)^2 \geq 0$$

(рівність нулю лише, коли $R = B$)

Таким чином, оцінка по регресії завжди точніша, ніж оцінка по відношенню чи середнє простої випадкової вибірки. Звичайно при цьому протрібно знати B.

4.4. Оцінювання по регресії при стратифікованому відборі: роздільні та сумісні оцінки

Знову, як і у параграфі 4.2 можна побудувати роздільні та сумісні оцінки по регресії при стратифікованому відборі. Скористаємося позначеннями параграфу 4.2.

Отже, роздільна оцінка по регресії така:

$$\hat{\mu}_{l_s} = \sum_{k=1}^{L} W_k \hat{\mu}_{l_k}, \text{ де } \hat{\mu}_{l_k} = \bar{y}_k + B_k (\mu_{x_k} - \bar{x}_k),$$

де B_k – коефіцієнт лінійної регресії k –ї страти

$$B_k = \frac{S_{yx_k}}{S^2_{x_k}} = \frac{\sum_{i=1}^{N_k}\left(Y_{ki} - \mu_{y_k}\right)\left(X_{ki} - \mu_{x_k}\right)\Big/(N_k - 1)}{\sum_{i=1}^{N_k}\left(X_{ki} - \mu_{x_k}\right)^2\Big/(N_k - 1)}$$

Тобто, в кожній страті підраховується своя оцінка по регресії:

Оцінка цього коефіцієнта

$$\hat{B}_k = \frac{\sum_{i=1}^{n_k}\left(y_{ki} - \bar{y}_k\right)\left(x_{ki} - \bar{x}_k\right)}{\sum_{i=1}^{n_k}\left(x_{ki} - \bar{x}_k\right)^2}.$$

$$\hat{\tau}_{l_s} = N\hat{\mu}_{l_s} = \sum_{k=1}^{L} N_k \hat{\mu}_{l_k}.$$

Ці оцінки будуть незміщеними і

$$D\hat{\mu}_{l_s} = \sum_{k=1}^{L}\frac{(1-f_k)}{n_k}\cdot W_k^2 \cdot S^2_{y_k}\left(1-\rho_k^2\right),\quad D\hat{\tau}_{l_s} = N^2 D\hat{\mu}_{l_s},$$

де ρ_k-коефіцієнт кореляції у k-й страті.

$$\rho_k = \frac{S_{yx_k}}{S_{x_k}S_{y_k}},\quad f_k = \frac{n_k}{N_k}.$$

Тепер побудуємо сумісну оцінку по регресії:

$$\hat{\mu}_{l_c} = \bar{y}_{st} + b\left(\mu_x - \bar{x}_{st}\right),$$

тобто лінійна регресія будується одна, на основі стратифікованих середніх по Y та X. Знайдемо b.

Тоді $D\hat{\mu}_{l_c} = \sum_{k=1}^{L} W_k^2 \frac{(1-f_k)}{n_k}\left(S^2_{y_k} - 2bS_{yx_k} + b^2 S^2_{x_k}\right),$

$$D\hat{\mu}_{l_c} \to min \text{ при} b = B_c = \frac{\sum_{k=1}^{L}\frac{W_k^2(1-f_k)}{n_k}S_{yx_k}}{\sum_{k=1}^{L}\frac{W_k^2(1-f_k)}{n_k}S^2_{x_k}} = \frac{\sum_{k=1}^{L}a_k B_k}{\sum_{k=1}^{L}a_k},$$

де $a_k = \frac{W_k^2(1-f_k)}{n_k}S^2_{x_k},\ B_k = \frac{S_{yx_k}}{S^2_{x_k}}.$

Тобто B_c - середнє зважене коефіцієнтів лінійної регресії в кожній страті.

$$\hat{\tau}_{l_c} = N\cdot\hat{\mu}_{l_c},\quad D\hat{\tau}_{l_c} = N^2\cdot D\hat{\mu}_{l_c}.$$

Підрахуємо тепер різницю дисперсій

$V = D\hat{\mu}_{l_c} - D\hat{\mu}_{l_s}$. Маємо

$$V = \sum_{k=1}^{L}\left[\frac{a_k}{S_{x_k}^2}\cdot S_{y_k}^2 - \frac{a_k}{S_{x_k}^2}\cdot 2B_c S_{yx_k} + B_c^2\cdot\frac{a_k}{S_{x_k}^2}\cdot S_{x_k}^2 - \frac{a_k}{S_{x_k}^2}\cdot S_{y_k}^2 + \frac{a_k}{S_{x_k}^2}\cdot\frac{S_{yx_k}^2}{S_{x_k}^2}\right] =$$

$$= \sum_{k=1}^{L} a_k (B_c - B_k)^2 \geq 0.$$

Тут була використана рівність

$$\frac{W_k(1-f_k)}{n_k} = \frac{a_k}{S_{x_k}^2}.$$

Отже, оптимальна роздільна оцінка по регресії має меншу дисперсію, ніж оптимальна сумісна оцінка. Якщо ж, всі B_k однакові, то ці дисперсії будуть співпадати. Але для оптимальних оцінок необхідно знати S_{xy_k} та $S_{x_k}^2$, $k = \overline{1, L}$.

Оцінки дисперсії при великих n_k

$$\upsilon(\hat{\mu}_{l_s}) = \sum_{k=1}^{L}\frac{W_k^2(1-f_k)}{n_k}\cdot\frac{1}{n_k-2}\left[\sum_{i=1}^{n_k}(y_{ki}-\bar{y}_k)^2 - \hat{B}_k\cdot\sum_{i=1}^{n_k}(x_{ki}-\bar{x}_k)^2\right].$$

Для оцінки B_c використовується статистика

$$\hat{B}_c = \sum_{k=1}^{L}\frac{W_k^2(1-f_k)}{n_k(n_k-1)}\sum_{k=1}^{n_k}(y_{ki}-\bar{y})(x_{ki}-\bar{x}) \Big/ \sum_{k=1}^{L}\frac{W_k^2(1-f_k)}{n_k(n_k-1)}\sum_{k=1}^{n_k}(x_{ki}-\bar{x})^2.$$

При пропорційному стратифікованому розміщенні $\frac{n_k}{N_k} = \frac{n}{N}$ для всіх $k = \overline{1, L}$ і при великих n_k, так що $n_k - 1 \approx n_k$

$$\hat{B}_c \approx \sum_{k=1}^{L}\sum_{i=1}^{n_k}(y_{ki}-\bar{y}_k)(x_{ki}-\bar{x}_k) \Big/ \sum_{k=1}^{L}\sum_{i=1}^{n_k}(x_{ki}-\bar{x}_k)^2$$ - МНК – оцінка.

$$\upsilon(\hat{\mu}_{l_c}) = \sum_{k=1}^{L}\frac{W_k^2(1-f_k)}{n_k(n_k-1)}\cdot\sum_{i=1}^{n_k}\left[(y_{ki}-\bar{y}_k) - \hat{B}_c(x_{ki}-\bar{x}_k)\right]^2.$$

Які краще використовувати оцінки, коли S_{xy_k} та $S_{x_k}^2$ невідомі? Зауважимо, що роздільні оцінки більш підвержені зсуву при невеликих обсягах вибірки всередині кожної страти, і, крім того, похибки при оцінці B_k складають значну долю дисперсії.

У сумісних оцінках значно збільшується похибка (дисперсія), якщо коефіцієнти регресії суттєво відрізняються від страти до страти. Якщо ж регресія лінійна і B_k однакові для всіх страт, то краще використовувати сумісні оцінки. Якщо ж регресія близька до лінійної і B_k

змінюються від страти до страти, то краще використовувати роздільні оцінки. Нарешті, якщо регресія нелінійна, а будемо використовувати лінійні оцінки, і обсяги вибірок невеликі, то краще використовувати сумісні оцінки.

4.5. Вправи

1. Із сукупності розподілу $N = 8$ взято просту випадкову вибірку обсягу $n = 2$, що дала такі результати для досліджуємої величини Y та допоміжної величини X : $y_1 = 50,\ x_1 = 10;\ y_2 = 22,\ x_2 = 2,\ \mu_x = 5$.

а) Знайти оцінку по відношенню середнього значення Y;

б) Оцінити також дисперсію $D\hat{\mu}_R$;

в) Побудувати 0,95-надійний інтервал для μ_y.

2. Рекламна фірма вивчає ефект застосування нової регіональної рекламної кампанії по продажу кави “Nestle”. Зроблена проста випадкова вибірка $n = 10$ магазинів з $N = 450$ регіональних магазинів, у яких продавалася кава. Зафіксована кількість проданих банок кави за місяць, протягом якого йшла реклама, а також кількість проданих банок кави у попередній до початку реклами місяць.

Магазин	Минулий місяць (x_i)	Даний місяць (y_i)
1	208	239
2	400	428
3	440	472
4	259	276
5	351	363
6	880	942
7	273	294
8	487	514
9	183	195
10	863	897

а) Використовуючи оцінку по відношенню оцінити τ_y, якщо $\tau_x = 216256$.

б) Оцінити τ_y за допомогою оцінки по регресії.

3. Сукупність з $N = 100$ одиниць розбита на дві страти, $N_1 = 30;\ N_2 = 70$. Вибірка обсягом $n = 8, n_1 = 4, n_2 = 4$ дала наступні результати для головної змінної Y та допоміжної X :

Страта 1		Страта 2	
x_{1i}	y_{1i}	x_{2i}	y_{2i}
2	0	10	7
5	3	18	15
9	7	21	10
15	10	25	16

а) Оцінити μ_y за допомогою сумісної та роздільної оцінок по відношенню, якщо $\mu_{x_1} = 8;\ \ \mu_{x_2} = 20$;

б) Оцінити μ_y за допомогою сумісної та роздільної оцінок по регресії.

4. Фермер бажає оцінити врожай яблук у саду, де росте $N = 200$ дерев. У минулому році він зібрав урожай $\tau_x = 5800$ кг. Проста випадкова вибірка десяти дерев дала такі результати

	1	2	3	4	5	6	7	8	9	10
x_i	30	21	25	29	34	23	19	29	35	26
y_i	29	23	26	30	34	24	22	30	38	29

а) Оцінити загальний урожай яблук τ_y, використовуючи оцінку по регресії та оцінку по відношенню.

б) Оцінити дисперсії цих оцінок.

//

(відповідь потрібно вписати у праву колонку)

Задача	Відповідь
1.Зроблена проста випадкова вибірка обсягу $n=10$із сукупності з 200 домогосподарств. Число людей у вибраних домогосподарствах становить: 3, 4, 5, 1, 3, 3, 5, 2, 2, 3. а) Оцінити загальне число людей в усіх домогосподарствах. Оцінити також дисперсію цієї оцінки.	
2.Для даних вправи 1 побудувати 0,95-надійний інтервал для середнього числа людей в одному домогосподарстві.	

Задача	Відповідь
3. Необхідно оцінити число дерев у деякому регіоні. Область вивчення була розділена на 200 одиниць або квадратів. З попереднього досвіду відомо, що $S^2 \approx 25$. а) Необхідно знайти обсяг простої випадкової вибірки квадратів для оцінки загального числа дерев в регіоні з точністю до 100 дерев і надійною ймовірністю 0,95.	
4.Обробляється2000 анкет. Проста випадковавибірка150 анкет при перевірці показала, що 38 анкет заповнені з помилками. Оцінитипропорціюпомилкових анкет у всійсукупності, а такожпобудувати 0,95-надійнийінтервал для цієїпропорції.	
5.Яким повинен бути розмір вибірки, щоб оцінити пропорцію людей з першою групою крові серед населення у 1500 людей з рівнем точності 0,02 та надійною ймовірністю 0,95? Відносноістиногозначенняпропорціїнемає єніякихпопередніхзнань.	