КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

О. І. ВАСИЛИК
Т. О. ЯКОВЕНКО

ЛЕКЦІЇ З ТЕОРІЇ І МЕТОДІВ ВИБІРКОВИХ ОБСТЕЖЕНЬ

Навчальний посібник

Рекомендовано Міністерством освіти і науки України
як навчальний посібник для студентів-магістрів
університетів, які навчаються за спеціальністю
«Статистика»

КИЇВСЬКИЙ УНІВЕРСИТЕТ 2010

УДК 512.642(075.8)
ББК 22.143я7
В19

Рецензенти:
д-р фіз.-мат. наук, проф. П. С. Кнопов,
д-р фіз.-мат. наук, проф. Р. Є. Майборода,
д-р екон. наук М. В. Пугачова

Затверджено вченою радою механіко-математичного факультету (протокол № 2 від 13 жовтня 2008 року)

Василик, О. І.

В19 Лекції з теорії і методів вибіркових обстежень : навчальний посібник / О. І. Василик, Т. О. Яковенко. – К. : Видавничо-поліграфічний центр "Київський університет", 2010. – 208 с.
ISBN 978-966-439-307-9

Посібник містить основні поняття та результати, які використовуються в теорії вибіркових обстежень, детальну інформацію про різні методи відбору елементів із генеральної сукупності та відповідні методи оцінювання її параметрів. Наводяться методи оцінювання лінійних і нелінійних функцій від сумарних значень кількох досліджуваних змінних, методи оцінювання за різницею, регресією та за відношенням. Розглядаються механізми породження пропусків, методи їх заповнення, оцінювання вибіркової дисперсії за даними із пропусками, аналіз даних із пропусками методом максимальної вірогідності.

Для студентів-магістрів спеціальності "Статистика" механіко-математичного факультету. Може бути корисним і для аспірантів, а також для тих, хто на практиці використовує вибірковий метод у процесі проведення обстежень.

УДК 512.642(075.8)
ББК 22.143я7

Гриф надано Міністерством освіти і науки України (лист № 1.11-7027 від 29.07.10)

ISBN 978-966-439-307-9

Зміст

Передмова

Посібник містить матеріал семестрових лекцій з нормативного курсу «Вибіркові обстеження», призначеного для студентів спеціальності «Статистика» механіко-математичного факультету Київського національного університету імені Тараса Шевченка.

Курс «Вибіркові обстеження» викладається на механіко-математичному факультеті КНУ близько десяти років. Спочатку це був 36-годинний спеціальний курс, в якому теоретичний матеріал супроводжувався розв'язанням практичних завдань і який ґрунтувався на книгах В. М. Пархоменко [5] та О. І. Черняка [7]. З перетворенням цього курсу в нормативний та збільшенням кількості аудиторних годин вдвічі (36 годин лекцій та 36 годин практичних занять) виникла потреба в додатковому теоретичному матеріалі. Упродовж кількох років ми використовували для лекцій з теорії та методів вибіркових обстежень багато джерел, основними з яких є [9, 10, 18, 20]. У результаті виникла ідея підготувати навчальний посібник, який містив би матеріали цих лекцій.

Необхідною умовою успішного засвоєння вмісту посібника є володіння базовими знаннями з теорії ймовірностей та математичної статистики.

Посібник складається з одинадцяти розділів, двох додатків та списку рекомендованої літератури. Перший розділ містить основні поняття, які використовуються в теорії вибіркових обстежень, а також означення та властивості π-оцінки Горвіца–Томпсона сумарного значення досліджуваної характеристики генеральної сукупності в загальному випадку. Застосування π-оцінки Горвіца–Томпсона та її властивостей для оцінювання параметрів генеральної сукупності при різних методах відбору розглядається в подальших розділах цього посібника. А саме, розділи з другого по восьмий містять інформацію про основні методи відбору елементів з генеральної сукупності та відповідні методи оцінювання параметрів генеральної сукупності. Зокрема, другий розділ присвячено простому випадковому відбору без повернення (ПВВбП), його основним властивостям, оцінюванню параметрів генеральної сукупності та підсукупностей, побудові довірчих інтервалів та

задачі визначення необхідного розміру вибірки у випадку застосування цього відбору. У третьому розділі наведено основні властивості та методи оцінювання для відбору Бернуллі. Тут вперше розглянуто поняття дизайн-ефекту, за допомогою якого визначається ефективність довільного методу відбору у порівнянні з ПВВбП. Четвертий розділ присвячений основним поняттям та результатам, які стосуються систематичного відбору. У п'ятому розділі розглядається відбір з поверненням, визначається оцінка Хансена–Гурвіца та досліджуються її властивості, а в шостому наведено деякі методи нерівноймовірнісного відбору без повернення. У сьомому розділі детально вивчається стратифікований відбір. Значна увага приділяється оптимальному та альтернативним розміщенням стратифікованої вибірки. Порівнюється ефективність цього вибіркового дизайну при оптимальному та пропорційному розміщеннях. Процедури одностадійного, двостадійного та багатостадійного кластерного відбору описані у восьмому розділі, де розглядаються як загальні випадки, так і випадок застосування ПВВбП на окремих стадіях одно- та двостадійного відбору. Наступні два розділи присвячені більш складним задачам оцінювання: в дев'ятому розділі розглядається оцінювання лінійних та нелінійних функцій від сумарних значень кількох досліджуваних змінних, а в десятому – оцінювання за різницею, за регресією та за відношенням. Одинадцятий розділ присвячено аналізу вибіркових даних за наявності пропусків (невідповідей). Розглядаються механізми породження пропусків, методи їх заповнення, оцінювання вибіркової дисперсії за даними з пропусками, аналіз даних з пропусками методом максимальної вірогідності та ін. Теоретичний матеріал супроводжується ілюстративними прикладами та вправами для самостійного розв'язання.

Посібник може бути корисним не тільки для студентів та аспірантів, які спеціалізуються з теорії ймовірностей та математичної статистики, а й для тих, хто на практиці використовує вибірковий метод при проведенні обстежень.

Автори підручника висловлюють щиру подяку рецензентам посібника – професору П. С. Кнопову, професору Р. Є. Майбороді та доктору економічних наук М. В. Пугачовій за цікаві дискусії, критичні зауваження та цінні поради щодо змісту цієї книги.

Ми дуже вдячні співробітникам кафедри теорії ймовірностей, статистики та актуарної математики Київського національного університету імені Тараса Шевченка, рідним та друзям за допомогу та підтримку під час написання книги. Також особливу подяку хочемо висловити професору Гуннару Кулдорфу з університету міста Умео (Швеція), за підтримки якого був розроблений та впроваджений курс «Вибіркові обстеження» на механіко-математичному факультеті нашого університету.

Вступ

Статистичне обстеження – це дослідження деякої підмножини або всієї сукупності предметів, осіб, подій, тощо з метою визначення їх кількісних параметрів: середнього або сумарного значення деякої характеристики, кількості або частки елементів з певною ознакою та ін. Прикладом статистичного обстеження є перепис населення країни. Він передбачає можливість 100-відсоткового обстеження всього населення.

Вибіркове обстеження полягає в аналізі всієї сукупності на основі інформації, отриманої за результатами обстеження лише частини сукупності – *вибірки*.

Інтенсивний розвиток теорії і методів вибіркових обстежень розпочався у 30-х роках XX сторіччя, коли у США та Європі виник великий попит на соціальну та економічну інформацію. Починаючи з 40-х років, вибіркові методи почали використовувати в переписах населення для отримання додаткової інформації, а з 50-х років в ООН почали широко застосовувати та пропагувати вибіркові обстеження з метою стимулювання різних країн до покращення методів збору та аналізу соціальної та економічної інформації. В Україні методи вибіркових обстежень почали широко застосовувати тільки після здобуття незалежності.

Використання вибіркових методів в обстеженнях дає помітну економію часу і коштів порівняно з суцільними обстеженнями, забезпечуючи при цьому потрібну точність результатів. У деяких випадках вибіркове обстеження є єдиним методом отримання необхідної інформації (наприклад, контроль якості, пов'язаний з руйнуванням зразків).

Сфера застосування вибіркових обстежень надзвичайно широка, але найбільше вони застосовуються в економіці, соціології, політології та психології. У рамках економічних та соціальних обстежень збирається інформація в країні та її областях про працездатне населення, соціальний добробут громадян, ціни та споживання, доходи і витрати домогосподарств, ринкові пріоритети, громадську думку, промислове і сільськогосподарське виробництво, навколишнє середовище тощо.

Множина елементів, для якої проводиться обстеження та відносно якої робляться висновки, називається *генеральною сукупністю* або *популяцією* (англ. `population`). Генеральні сукупності бувають скінченними і нескінченними. Нескінченні генеральні сукупності виникають у фізико-технічних, хімічних та інших природничих дослідженнях. Ми будемо мати справу тільки зі скінченними генеральними сукупностями. Кожному елементу генеральної сукупності відповідає значення однієї або кількох *досліджуваних характеристик* (*змінних*). Метою обстеження є отримання інформації про невідомі *параметри генеральної сукупності*, які є функціями від досліджуваних змінних.

Множину елементів, з якої практично проводиться відбір, називають *вибірковою сукупністю* (англ. `sampled population`). Часто генеральна сукупність та вибіркова сукупність співпадають, але в деяких випадках вони можуть і відрізнятись. Наприклад, коли при проведенні обстеження домогосподарств важко ідентифікувати всі домогосподарства в деякому місті, але доступна інформація про всіх, хто мешкає в цьому місті, та можна визначити до якого домогосподарства належить особа, тоді можна створити вибірку осіб і обстежувати ті домогосподарства, до яких належать вибрані особи. У цьому випадку генеральною сукупністю буде сукупність усіх домогосподарств, а вибірковою сукупністю буде сукупність осіб, що проживають в цьому місті. У цьому навчальному посібнику ми будемо вважати для простоти, що ці дві сукупності є ідентичними.

Зазвичай для проведення вибіркового обстеження деякої сукупності потрібно мати реєстр (список) її елементів. Такий реєстр називається *вибірковою основою* (англ. `sampling frame`). Вибіркова основа дає можливість встановити зв'язок між елементами генеральної сукупності та *вибірковими одиницями* (англ. `sampling units`). Залежно від рівня складності обстеження вибіркові одиниці можуть бути як окремими елементами сукупності, так і групами елементів. Із генеральної сукупності відбирається деяка частина (підмножина) елементів, яка називається *вибіркою* (англ. `sample`). Кількість елементів у ній називається *розміром* (*обся-*

гом, об'ємом) вибірки. Вибірка називається *ймовірнісною* (*випадковою*), якщо вона одержана за допомогою деякого ймовірнісного правила відбору. Іноді вибірки формуються не випадковим чином, а згідно з «експертним рішенням» (тобто певна уповноважена особа–«експерт» вирішує, які саме елементи повинні потрапити до вибірки) або за згодою осіб, стосовно яких проводиться обстеження (вони самі визначають, брати участь в обстеженні чи ні). Це приклади «невипадкових» вибірок. За такими вибірками можна оцінювати деякі параметри генеральної сукупності, але, на жаль, не можна нічого сказати про точність результуючих оцінок, оскільки неможливо застосувати при цьому ймовірнісний підхід. У цьому посібнику ми будемо розглядати тільки ймовірнісні вибірки.

На етапі планування будь-якого обстеження необхідно знайти відповіді на такі типові запитання:

- що є метою обстеження; яку інформацію необхідно одержати і які характеристики треба виміряти і обчислити;
- що являє собою генеральна сукупність та вибіркова сукупність; що буде елементом обстеження (адміністративний район, підприємство, сім'я, окрема особа тощо);
- яким чином буде отримано необхідну інформацію (за допомогою наявної статистичної звітності, інтерв'ю, анкетуванням по пошті чи по телефону, простим спостереженням тощо);
- яким має бути необхідний рівень точності висновків, характер допустимого ризику від помилки та розмір збитків у разі прийняття помилкового рішення;
- якими повинні бути фінансові та інші ресурси, вартість кожної операції;
- якою має бути кваліфікація та рівень підготовки персоналу;
- яке обстеження потрібно провести: суцільне чи вибіркове;

- у разі проведення вибіркового обстеження – яким має бути метод відбору та розмір вибірки, щоб гарантувати необхідну точність;
- які методи оцінювання параметрів генеральної сукупності та перевірки статистичних гіпотез будуть застосовані;
- в якому вигляді будуть подані результати обстеження (звіти, публікації); яким буде ступінь деталізації та секретності інформації.

У разі проведення вибіркового обстеження після формування вибірки обчислюють значення досліджуваних характеристик для всіх елементів, що потрапили у вибірку. На основі отриманих вибіркових даних обчислюють значення оцінок досліджуваних параметрів генеральної сукупності, оцінюють точність отриманих результатів.

Результуючі оцінки містять похибки, які можна поділити на два типи відповідно до джерела їх виникнення:

- *вибіркові похибки* виникають у результаті того, що замість усієї генеральної сукупності досліджується тільки її частина;
- *невибіркові похибки* виникають при вимірюванні та збереженні інформації, внаслідок відсутності відповіді (пропуски у вибіркових даних) або її неправдивості тощо.

Вибіркові обстеження можуть дуже відрізнятися за рівнем складності. Це залежить як від структури і розміру генеральної сукупності, так і від кількості характеристик, які потрібно дослідити. Наприклад, при проведенні соціологічного обстеження дослідника можуть цікавити відповіді на декілька запитань, а при проведенні обстеження видатків домогосподарств можуть досліджуватися більше сотні характеристик (змінних).

У цьому посібнику ми зосередимося на теоретико-ймовірнісних та статистичних аспектах планування вибіркового обстеження та

аналізу вибіркових даних. У всіх розділах, крім останнього, будемо розглядати дещо ідеальну ситуацію, коли всі невибіркові похибки відсутні. Основна частина цієї книги присвячена вивченню методів оцінювання параметрів генеральної сукупності *на основі вибіркового дизайну* (англ. `design-based approach`). Це означає, що на момент проведення обстеження значення досліджуваної характеристики для окремого елемента генеральної сукупності є сталим, а варіація значень оцінок параметрів виникає внаслідок випадкового методу відбору елементів до вибірки. Метод аналізу даних з пропусками за допомогою функції вірогідності, який розглядається в останньому розділі посібника, є прикладом застосування іншого підходу – оцінювання на основі моделі (англ. `model-based` або `model-dependent approach`). Методи оцінювання на основі моделей ґрунтуються на припущенні, що структура генеральної сукупності відповідає певній моделі «суперпопуляції». У моделі суперпопуляції задається N-вимірний розподіл випадкового вектора $\mathbf{Y} = (Y_1, \ldots, Y_N)$, де Y_k – випадкова величина, яка описує значення досліджуваної характеристики y для k-го елемента генеральної сукупності. Справжній вектор значень характеристики y для елементів генеральної сукупності, тобто $\mathbf{y} = (y_1, \ldots, y_N)$, розглядається як реалізація випадкового вектора $\mathbf{Y}$. Оскільки у цьому випадку для побудови оцінок невідомих параметрів генеральної сукупності використовуються припущення про розподіл вектора $\mathbf{Y}$, то якість цих оцінок залежить від того, наскільки вдало підібрана модель. Також у посібнику наведено деякі методи оцінювання з використанням допоміжної інформації, що є прикладом *оцінювання за допомогою моделей* (англ. `model-assisted approach`). Цей підхід полягає в комбінації методів оцінювання на основі дизайну та на основі моделей і дає можливість використовувати переваги цих методів.

Розділ 1

Основні поняття

1.1. Генеральна сукупність, вибірка, схеми відбору

Розглянемо генеральну сукупність, що складається з N пронумерованих елементів

$$U = \{u_1, u_2, ..., u_k, ..., u_N\}.$$

Для простоти будемо задавати елемент його порядковим номером

$$U = \{1, 2, ..., k, ..., N\}.$$

Будемо вважати, що кількість елементів у генеральній сукупності відома, хоча на практиці при проведенні обстежень N може бути невідомою величиною.

Нехай змінна y – це характеристика генеральної сукупності, яка нас цікавить, а y_k – значення цієї характеристики для k-го елемента сукупності. Значення y_k, $k \in U$, невідомі до проведення обстеження. Нехай потрібно оцінити сумарне або середнє значення характеристики y:

$$T = \sum_{k \in U} y_k \qquad \text{та} \qquad \overline{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{T}{N}.$$

Величини: сумарне T, середнє $\overline{Y}$, дисперсія генеральної сукупності S^2 є найпростішими прикладами параметрів, що характеризують генеральну сукупність. Ми можемо отримати значення цих параметрів, якщо обстежимо всю генеральну сукупність, або можемо оцінити їх, обстеживши лише частину сукупності, що буде значно дешевше та займе менше часу. Спочатку відбирається частина генеральної сукупності – вибірка, а потім обстежуються елементи, що потрапили у вибірку. На основі отриманих значень обчислюються оцінки невідомих параметрів сукупності.

Зауважимо, що для оцінювання деякого параметра генеральної сукупності можуть бути використані різні оцінки (методи оцінювання). Наприклад, для оцінювання середнього використовують такі оцінки: середнє арифметичне, середнє геометричне, мода, медіана, обчислені за вибіркою. Тобто, середнє значення досліджуваної характеристики генеральної сукупності є параметром, а вибіркове середнє арифметичне, вибіркова мода є його оцінками. Значення оцінок змінюються залежно від того, які елементи генеральної сукупності потрапляють у вибірку[1].

Позначимо вибірку літерою s (від англ. `sample`). У загальному випадку вибіркою може бути будь-яка підмножина генеральної сукупності. Будемо розглядати лише ймовірнісні вибірки, тобто такі, що були отримані за допомогою деякої ймовірнісної схеми відбору. Ймовірнісна вибірка формується в результаті проведення серії випадкових випробувань.

Існує два основні види ймовірнісних схем відбору – схеми на основі *жеребкування* та схеми на основі *послідовного відбору зі списку*. Схеми відбору на основі *жеребкування* полягають у проведенні серії випадкових випробувань, у результаті кожного з яких один елемент генеральної сукупності або її частини потрапляє у вибірку.

Приклад 1.1. **Простий випадковий відбір без повернення.**

З генеральної сукупності розміру N відбирається n елементів без повернення ($n \leq N$). Для цього потрібно провести n випадкових випробувань.

1) Випадковим чином вибирається один елемент з N елементів генеральної сукупності. Ймовірність бути вибраним однакова для всіх елементів генеральної сукупності і дорівнює $\frac{1}{N}$.

2) Випадковим чином вибирається другий елемент з $N-1$ тих, що залишились. Імовірність потрапити до вибірки однакова для всіх цих елементів і дорівнює $\frac{1}{N-1}$.

...

n) Вибирається n-й елемент з $N-n+1$ елементів, що залишились. Імовірність бути вибраним на цьому кроці однакова для всіх $N-n+1$ елементів і дорівнює $\frac{1}{N-n+1}$.

У результаті отримаємо ймовірнісну вибірку розміру n. ◊

[1] Дуже важливо розрізняти поняття «оцінки» (англ. `estimator`) – деякої функції від вибірки, за допомогою якої можна оцінити потрібний параметр, та «значення оцінки» (англ. `estimate`) – конкретного числа, яке є результатом застосування цієї функції до конкретних вибіркових даних.

За схемою на основі *послідовного відбору зі списку*: для кожного елемента, починаючи з початку списку (але не обов'язково до кінця), проводиться випадкове випробування, результатом якого є включення або, навпаки, невключення елемента, що розглядається, у вибірку.

Приклад 1.2. **Відбір Бернуллі.** Розглянемо впорядковану в список генеральну сукупність $U = \{1, 2, ..., N\}$. Нехай наперед задано деяке число π таке, що $0 < \pi < 1$, та набір N незалежних реалізацій рівномірно розподіленої на $[0, 1]$ випадкової величини: $\varepsilon_1, \varepsilon_2, ..., \varepsilon_N$. Кожному елементу k ставиться у відповідність значення ε_k. Якщо $\varepsilon_k < \pi$, то цей елемент відбирається, в іншому випадку – ні. Зрозуміло, що при такій схемі відбору, ймовірність того, що елемент буде відібраний, дорівнює π для кожного з N елементів, та при $k \neq l$ події «k-й елемент вибраний» та «l-й елемент вибраний» є незалежними. Кількість елементів, що потрапили таким чином у вибірку, є біноміально розподіленою випадковою величиною з параметрами N та π. ◊

1.2. Вибірковий дизайн

Знаючи, яким чином проводиться відбір у вибірку, завжди можна (хоча не завжди просто) підрахувати ймовірність отримання конкретної вибірки s. Позначимо цю ймовірність через $p(s)$. Отже, функція $p(\cdot)$ підраховує ймовірність отримання вибірки s при заданій схемі відбору. Ця функція відіграє центральну роль у теорії вибіркових обстежень, оскільки визначає дуже важливі

статистичні характеристики: ймовірнісний розподіл, математичне сподівання та дисперсію оцінюючих статистик, що обчислюються на основі інформації, отриманої з вибірки. У зарубіжній літературі цю функцію називають *вибірковий дизайн* (англ. `sampling design`) [20].

Приклад 1.3. Для вибіркової схеми, що була наведена в прикладі 1.1 і яка відповідає *простому випадковому відбору без повернення*, з генеральної сукупності N елементів можна утворити C_N^n різних вибірок розміру n. При цьому всі вибірки рівноможливі. Тому ймовірність того, що з них буде обрана одна конкретна вибірка, дорівнює $1/C_N^n$. Отже, в цьому випадку вибірковий дизайн має вигляд

$$p(s) = \begin{cases} 1/C_N^n, & \text{якщо вибірка } s \text{ складається з } n \text{ елементів;} \\ 0, & \text{в іншому випадку.} \end{cases}$$

◊

Будемо позначати вибірковий дизайн, який відповідає простому випадковому відбору без повернення, абревіатурою ПВВбП (англ. `simple random sampling without replacement`).

Розглянемо випадковий об'єкт $\mathcal{S}$, можливими значеннями якого є вибірки s, що обумовлені вибірковим дизайном $p(\cdot)$. Нехай $\mathfrak{F}$ – σ-алгебра всіх підмножин (вибірок), які можна утворити з елементів генеральної сукупності U. До $\mathfrak{F}$ належать всі 2^N підмножини сукупності U, і в тому числі $\varnothing$ та U. Тоді

$$P(\mathcal{S} = s) = p(s) \quad \forall s \in \mathfrak{F}.$$

Функція $p(\cdot)$ задає ймовірнісний розподіл на множині $\mathfrak{F}$, а тому:

1) $p(s) \geq 0$ для всіх $s \in \mathfrak{F}$;

2) $\sum\limits_{s \in \mathfrak{F}} p(s) = 1$.

Зауважимо, що деякі вибірки з множини $\mathfrak{F}$ матимуть нульову ймовірність бути обраними. Підмножина множини $\mathfrak{F}$, що складається з тих вибірок s, що мають ненульову ймовірність бути обраними, є *множиною можливих вибірок*. Лише вони можуть бути вибраними в рамках застосування певного вибіркового дизайну.

Розмір вибірки n_s дорівнює кількості елементів сукупності, що потрапили у вибірку s. Ця величина не обов'язково повинна бути сталою, як у випадку простого випадкового відбору без повернення з прикладу 1.1 ($n_s = n$ для всіх можливих вибірок s). При застосуванні відбору Бернуллі розміри можливих вибірок відрізняються.

Приклад 1.4. Яким же буде вибірковий дизайн при застосуванні відбору Бернуллі? Нехай n_s – розмір вибірки s. Ймовірність того, що елемент потрапляє у вибірку, дорівнює π та випробування при застосуванні *схеми послідовного відбору зі списку* незалежні між собою. Тому

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s}, \quad s \in \mathfrak{F}.$$

Ймовірність того, що розмір вибірки при відборі Бернуллі дорівнює m, становить

$$C_N^m \pi^m (1-\pi)^{N-m}, \quad m = 0, 1, 2, ..., N.$$

Таким чином, розмір вибірки при відборі Бернуллі – це біноміально розподілена випадкова величина з математичним сподіванням $N\pi$ та дисперсією $N\pi(1-\pi)$. ◊

Будемо позначати вибірковий дизайн, який відповідає відбору Бернуллі, через ВБ.

Вибірковий дизайн $p(s)$ визначає статистичні властивості оцінок, що використовують інформацію з вибірки. Але в основному – це тільки математичний інструмент, який рідко використовується на практиці.

Важливо зауважити, що різні схеми відбору можуть приводити до одного і того самого вибіркового дизайну.

Приклад 1.5. Дизайн, що відповідає простому випадковому відбору без повернення (ПВВбП), можна отримати за допомогою альтернативної схеми відбору.

Один елемент відбирається з імовірністю $\frac{1}{N}$ і повертається назад. Процедура повторюється, поки не отримаємо n різних елементів. Загальна кількість відборів ν, яку потрібно зробити, більша за n з імовірністю одиниця і має досить складний розподіл. Але ця процедура також визначає ПВВбП-дизайн. $\Diamond$

При проведенні обстеження основною метою є отримання оцінок для одного чи декількох параметрів генеральної сукупності (сумарного значення, середнього тощо). При цьому потрібно зробити вибір щодо двох важливих питань:

1) який вибірковий дизайн та яку відповідну йому схему відбору доцільно застосувати;

2) який метод використати для оцінювання досліджуваних параметрів генеральної сукупності.

Вибір по кожному з цих пунктів не є незалежним. Сукупність схеми відбору та методу оцінювання називається *стратегією*. Для досліджуваного параметра сукупності потрібно знайти найкращу можливу стратегію, при якій оцінки будуть найточнішими.

У наступних розділах ми розглянемо основні відбори. Спочатку дослідимо прості відбори. Відбір будемо вважати простим, якщо

1) для нього існує вибіркова основа, за якою можна ідентифікувати кожен елемент генеральної сукупності;

2) вибірковими одиницями є елементи генеральної сукупності.

До простих відборів належать:

- простий випадковий відбір без повернення (ПВВбП);
- відбір Бернуллі (ВБ);
- систематичний відбір (СВ);
- простий випадковий відбір з поверненням (ПВВзП);

- відбір Пуассона (ВП);
- відбір, пропорційний розміру: без повернення та з поверненням (ВПР);
- стратифікований відбір (СТВ).

В основному будемо розглядати сумарне значення T досліджуваної характеристики генеральної сукупності та незміщені оцінки $\widehat{t}$ цього параметра. Особливу увагу будемо приділяти оцінці Горвіца–Томпсона $\widehat{t}_\pi$ при відборі без повернення та оцінці Хансена–Гурвіца $\widehat{t}_{pwr}$ при відборі з поверненням, а також їх основним характеристикам: дисперсії та оцінці дисперсії.

Але перед цим нам потрібно ввести та дослідити деякі важливі поняття.

1.3. Поняття «статистика»

У теорії математичної статистики, під терміном «статистика» розуміють дійснозначну функцію $Q(\mathcal{S})$, значення якої змінюється залежно від результату деякого випадкового експерименту. Основною вимогою до статистики є те, щоб вона була визначена для будь-якого результату експерименту. Метою є встановлення, як змінюється статистика $Q(\mathcal{S})$ залежно від того, якого значення s буде набувати випадкова множина $\mathcal{S}$.

Наведемо приклади статистик:

- $Q_1(\mathcal{S}) = I_k(\mathcal{S})$, $k \in U$ – індикатор включення;
- $Q_2(\mathcal{S}) = n_{\mathcal{S}} = \sum_{k\in U} I_k(\mathcal{S})$ – розмір вибірки;
- $Q_3(\mathcal{S}) = \sum_{k\in\mathcal{S}} y_k$ – вибіркове сумарне значення змінної y;
- $Q_4(\mathcal{S}) = \frac{1}{n_S}\sum_{k\in\mathcal{S}} y_k$ – вибіркове середнє по змінній y;
- $Q_5(\mathcal{S}) = \sum_{k\in\mathcal{S}} y_k / \sum_S z_k$ – відношення двох вибіркових сумарних значень по змінних y та z.

Функція $\sum_{k\in\mathcal{S}} y_k / \sum_{k\in U} z_k$ не є статистикою, оскільки вона не визначена у випадку, коли немає інформації по всій генеральній сукупності для змінної z.

На практиці, коли вибірка обрана, ми отримуємо одну єдину реалізацію s випадкової множини $\mathcal{S}$. Для отриманої вибірки s ми будемо вважати, що можливо обстежити і виміряти змінні (y, z та інші) для кожного елемента $k \in s$. Наприклад, для статистики $Q_5(\mathcal{S}) = \sum_{k\in\mathcal{S}} y_k / \sum_{k\in\mathcal{S}} z_k$ ми можемо після вимірювання підрахувати її значення, а саме $Q_5(s) = \sum_{k\in s} y_k / \sum_{k\in s} z_k$.

Зауважимо, що при підході, який *базується на основі дизайну*, вважають, що змінні y та z хоч і набувають різних значень для різних елементів k, але вони не вважаються випадковими величинами. Випадкова природа статистики Q при цьому повністю пояснюється тим, що множина $\mathcal{S}$ є випадковою.

Для статистики $Q(\mathcal{S})$, як і для будь-якої випадкової величини, визначені основні статистичні характеристики – математичне сподівання та дисперсія:

$$\begin{aligned} E(Q) &= \sum_{s\in\mathfrak{F}} p(s)Q(s); \\ \mathcal{D}(Q) &= E[Q - E(Q)]^2 = \sum_{s\in\mathfrak{F}} p(s)[Q - E(Q)]^2, \end{aligned}$$

де $\mathfrak{F}$ – σ-алгебра всіх підмножин (вибірок), які можна утворити з елементів генеральної сукупності U.

Коваріація між двома статистиками $Q_1 = Q_1(\mathcal{S})$ та $Q_2 = Q_2(\mathcal{S})$ визначається так:

$$\begin{aligned} \operatorname{cov}[Q_1, Q_2] &= E[Q_1 - EQ_1][Q_2 - EQ_2] = \\ &= \sum_{s\in\mathfrak{F}} p(s)[Q_1(s) - E(Q_1(s))][Q_2(s) - E(Q_2(s))]. \end{aligned}$$

Для простих статистик $Q(\mathcal{S})$, наприклад *лінійних*, математичне сподівання та дисперсію можна досить легко обчислити за допомогою аналітичних формул. Більш загальним методом, за допомогою якого можна обчислити ці величини, є метод Монте-Карло. Він полягає в послідовному отриманні великої кількості

вибірок за допомогою однієї і тієї самої схеми відбору та підрахунку значення статистики $Q(\mathcal{S})$ для кожної отриманої вибірки. І потім на основі отриманих значень можна порахувати середнє і дисперсію цих значень. Вони будуть близькими до дійсних значень математичного сподівання та дисперсії цієї статистики, за умови що кількість згенерованих вибірок досить велика.

Зазвичай метод Монте-Карло використовується для дослідження властивостей складних статистик, але основним його недоліком, що унеможливлює його використання при вибіркових обстеженнях, є те, що на практиці отримують лише одну вибірку.

1.4. Оцінки та їх властивості

Більшість статистик, що розглядаються в теорії вибіркових обстежень, – це різні види оцінок. Оцінка – це статистичне знаряддя для підрахунку величин, що є близькими до деякої характеристики генеральної сукупності, яку ми хочемо оцінити, – параметра генеральної сукупності. Загальновживане позначення параметра сукупності – θ.

Якщо вивчається лише одна змінна y, то θ розглядається як функція від N значень змінної y: $y_1, y_s, ..., y_N$. Тобто

$$\theta = \theta(y_1, y_s, ..., y_N).$$

Приклад 1.6.

$\theta_1 = T = \sum_{k\in U} y_k$ – сумарне по генеральній сукупності;

$\theta_2 = \overline{Y} = \frac{1}{N}\sum_{k\in U} y_k$ – середнє по генеральній сукупності;

$\theta_3 = S_y^2 = \frac{\sum_{k\in U}(y_k - \overline{Y})^2}{N-1}$ – дисперсія генеральної сукупності;

$\theta_4 = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k}$ – відношення (функція декількох змінних).

◊

Оцінку параметра θ будемо позначати $\hat{\theta} = \hat{\theta}(\mathcal{S})$. Якщо s – це реалізація випадкової множини $\mathcal{S}$, то будемо припускати, що можливо підрахувати значення $\hat{\theta}$ за значеннями $y_k, z_k, ...$, що відповідають елементам $k \in s$.

Приклад 1.7.

$$\hat{\theta}_1 = \frac{N}{n}\sum_{k\in\mathcal{S}} y_k \text{ – оцінка параметра } \theta_1 = \sum_{k\in U} y_k;$$

$$\hat{\theta}_4 = \frac{\sum_{k\in\mathcal{S}} y_k}{\sum_{k\in\mathcal{S}} z_k} \text{ – оцінка параметра } \theta_4 = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k}.$$

◊

Означення 1.1. *Вибірковим розподілом* оцінки $\hat{\theta}$ є набір усіх можливих значень оцінки $\hat{\theta}$ та відповідні їм імовірності

$$P\{\hat{\theta} = c\} = \sum_{s:\hat{\theta}(s)=c} p(s). \tag{1.1}$$

Математичне сподівання та дисперсія оцінки визначаються відносно цього розподілу:

$$E(\hat{\theta}) = \sum_{s\in\mathfrak{F}} p(s)\hat{\theta}(s);$$

$$\mathcal{D}(\hat{\theta}) = \sum_{s\in\mathfrak{F}} p(s)[\hat{\theta}(s) - E(\hat{\theta})]^2.$$

Найважливішими характеристиками якості оцінки $\hat{\theta}$ є:

1) *Зміщення* (англ. `bias`)

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Оцінка для параметра θ називається *незміщеною*, якщо $B(\hat{\theta}) = 0$, що рівносильно $E(\hat{\theta}) = \theta$ для будь-якого набору $y_1, y_s, ..., y_N$.

2) *Середньоквадратична похибка* (англ. `mean square error`)

$$MSE(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = \sum_{s\in\mathfrak{F}} p(s)[\hat{\theta}(s) - \theta]^2.$$

Для MSE справедлива рівність $MSE(\hat{\theta}) = \mathcal{D}(\hat{\theta}) + [B(\hat{\theta})]^2$, звідки випливає, що для незміщених оцінок $MSE(\hat{\theta}) = \mathcal{D}(\hat{\theta})$.

3) *Середньоквадратичне відхилення* або *стандартна похибка* (англ. `standard error`) оцінки $\hat{\theta}$, що дорівнює квадратному кореню з дисперсії $\sigma = \sqrt{\mathcal{D}(\hat{\theta})}$.

4) *Відносна стандартна похибка* або *коефіцієнт варіації* (англ. `coefficient of variation`), що є відношенням стандартної похибки оцінки до математичного сподівання

$$CV(\hat{\theta}) = \frac{\sqrt{\mathcal{D}(\hat{\theta})}}{E(\hat{\theta})}.$$

Для оцінювання цього коефіцієнта часто використовують таку оцінку:

$$cv(\hat{\theta}) = \frac{\sqrt{\hat{\mathcal{D}}(\hat{\theta})}}{\hat{\theta}}.$$

Зауваження 1.1. Звернемо увагу на різницю між випадковою множиною $\mathcal{S}$ та вибіркою s. Отримані в результаті відбору вибірки s є можливими значеннями випадкової множини $\mathcal{S}$, аналогічно тому, як значення $x_1, x_2, ..., x_m$ є можливими значеннями деякої випадкової величини X. У подальшому при підрахунку статистик виду $\sum_{k\in\mathcal{S}} y_k$ не будемо наголошувати на тому, що її значення можна підрахувати для будь якого можливого значення $\mathcal{S}$ – вибірки s і будемо просто писати $\sum_{k\in s} y_k$.

1.5. Імовірності включення елемента у вибірку

Елементи генеральної сукупності не обов'язково мають однакову ймовірність бути вибраними. Деякі елементи можуть бути більш «важливими», тобто мати більшу ймовірність потрапити у вибірку, інші – меншу. За рахунок надання різним елементам різних імовірностей можна отримати точніші оцінки тих параметрів, які нас цікавлять.

Припустимо, що для обстеження обрано деякий вибірковий дизайн $\{p(s), s \in \mathfrak{F}\}$. Розглянемо випадкову величину, що є індикатором потрапляння елемента k у вибірку:

$$I_k = \begin{cases} 1, & \text{якщо } k \in \mathcal{S}; \\ 0, & \text{в іншому випадку.} \end{cases}$$

I_k є функцією від випадкового об'єкта $\mathcal{S}$: $I_k = I_k(\mathcal{S})$. Назвемо цю величину *індикатором включення*.

Ймовірність того, що елемент k включений у вибірку, називається ймовірністю включення першого порядку, позначається через π_k та обчислюється за формулою:

$$\pi_k = P(k \in \mathcal{S}) = P(I_k = 1) = \sum_{s \ni k} p(s).$$

Тут вираз $\sum\limits_{s \ni k}$ означає, що сума береться по всіх вибірках, які містять елемент k.

Ймовірність того, що елементи k та l одночасно включені у вибірку (будемо позначати це як $k\&l \in \mathcal{S}$), називається ймовірністю включення другого порядку, позначається через π_{kl} та обчислюється, за умови відомого вибіркового дизайну, так:

$$\pi_{kl} = P(k\&l \in \mathcal{S}) = P(I_k I_l = 1) = \sum_{s \ni k\&l} p(s).$$

При цьому $\pi_{kl} = \pi_{lk}$ для всіх $l, k = \overline{1, N}$. При $k = l$ будемо мати

$$\pi_{kk} = P(I_k^2 = 1) = P(I_k = 1) = \pi_k.$$

Кожному вибірковому дизайну відповідає набір N імовірностей включення першого порядку

$$\pi_1, \pi_2, ..., \pi_k, ..., \pi_N$$

та набір з $C_N^2 = N(N-1)/2$ різних імовірностей включення другого порядку

$$\pi_{12}, \pi_{13}, ..., \pi_{kl}, ..., \pi_{N-1,N}.$$

За потреби можна обчислити і ймовірності включення вищих порядків, але вони не є настільки важливими.

Запишемо ймовірності включення першого та другого порядків для двох вибіркових дизайнів – ПВВбП та ВБ.

Приклад 1.8. При ПВВбП кількість вибірок розміру n із сукупності N елементів, які містять елемент k, дорівнює C_{N-1}^{n-1}. А кількість таких вибірок, які містять обидва елементи k та l, дорівнює C_{N-2}^{n-2}. Всі такі вибірки мають однакову ймовірність бути вибраними: $p(s) = 1/C_N^n$. Отже,

$$\pi_k = P(I_k = 1) = \sum_{s \ni k} p(s) = C_{N-1}^{n-1} \cdot \frac{1}{C_N^n} = \frac{n}{N}, \quad k = \overline{1, N};$$

$$\pi_{kl} = P(I_k I_l = 1) = \sum_{s \ni k\&l} p(s) = C_{N-2}^{n-2} \cdot \frac{1}{C_N^n} = \frac{n(n-1)}{N(N-1)},$$

$$k \neq l = \overline{1, N}. \quad \Diamond$$

Приклад 1.9. При ВБ всі індикатори включення I_k – незалежні та однаково розподілені випадкові величини. Кожен елемент потрапляє у вибірку з імовірністю π. Отже,

$$\pi_k = P(I_k = 1) = \pi, \quad k = \overline{1, N};$$
$$\pi_{kl} = P(I_k I_l = 1) = P(I_k = 1) \cdot P(I_l = 1) = \pi^2, \quad k \neq l = \overline{1, N}.$$

$\Diamond$

Вибірковий дизайн часто обирається таким чином, щоб отримати бажані ймовірності включення першого і другого порядків. Навіть якщо сама функція $p(\cdot)$ має досить складний вигляд, для визначення математичного сподівання та дисперсії оцінок достатньо знати відповідні ймовірності включення π_k та π_{kl}.

Означення 1.2. Вибірковий дизайн називається *випадковим*, якщо всі елементи генеральної сукупності мають додатні ймовірності включення першого порядку: $\pi_k > 0$ для всіх $k \in U$, тобто кожен елемент має шанс потрапити у вибірку.

На практиці інколи використовують вибіркові дизайни, для яких ця умова не виконується. Наприклад, при деяких обстеженнях підприємств малі підприємства мало впливають на загальну картину, тому їх не розглядають, тобто вони мають нульові ймовірності включення першого порядку. Але потрібно дуже обережно до цього ставитися, оскільки виключення елементів з обстеження може призвести до суттєвого зміщення оцінок.

При простих схемах відбору ймовірності π_k відомі наперед, до початку відбору. Але при використанні більш складних, багатостадійних схем відбору ймовірності включення у вибірку не завжди відомі заздалегідь.

Означення 1.3. Вибірковий дизайн називається *вимірним*, якщо для всіх елементів сукупності ймовірності включення першого і другого порядків додатні. Тобто

$$\pi_k > 0 \quad \text{та} \quad \pi_{kl} > 0 \quad \forall k \neq l \in U.$$

Умова *вимірності* дизайну дозволяє отримувати обґрунтовані оцінки для дисперсій оцінок та будувати довірчі інтервали.

1.6. Індикатор включення елемента у вибірку

Дослідимо спочатку статистичні властивості найпростішої статистики – індикатора включення елемента у вибірку.

***Твердження* 1.1.** *Для будь-якого вибіркового дизайну $p(\cdot)$ та для всіх k, $l = 1, 2, ..., N$:*

$$\begin{aligned} E(I_k) &= \pi_k; \\ \mathcal{D}(I_k) &= \pi_k(1-\pi_k); \\ \operatorname{cov}(I_k, I_l) &= \pi_{kl} - \pi_k\pi_l. \end{aligned}$$

Доведення. Індикатор $I_k = I_k(\mathcal{S})$ набуває лише двох значень: 1 – з імовірністю π_k, тобто, коли елемент k потрапляє у вибірку, та 0 – з імовірністю $1-\pi_k$ у протилежному випадку. Отже, $E(I_k) = \pi_k$. Оскільки $E(I_k) = E(I_k^2) = \pi_k$, то $\mathcal{D}(I_k) = E(I_k^2) - (E(I_k))^2 = \pi_k - \pi_k^2 = \pi_k(1-\pi_k)$.

Крім того, $I_kI_l = 1$ лише тоді, коли обидва елементи k та l потрапили у вибірку. Отже, $E(I_kI_l) = P(I_kI_l = 1) = \pi_{kl}$, а тому $\operatorname{cov}(I_k, I_l) = E(I_kI_l) - E(I_k)E(I_l) = \pi_{kl} - \pi_k\pi_l$. □

Ще однією простою статистикою є *розмір вибірки* $n_{\mathcal{S}}$. Зобразимо випадкову величину $n_{\mathcal{S}}$ у вигляді суми індикаторів включення

$$n_{\mathcal{S}} = \sum_{k\in U} I_k,$$

звідки відразу отримаємо вирази для обчислення основних характеристик цієї статистики

$$E(n_{\mathcal{S}}) = \sum_{k\in U} E(I_k) = \sum_{k\in U} \pi_k; \tag{1.2}$$

$$\begin{aligned} \mathcal{D}(n_{\mathcal{S}}) &= \mathcal{D}\left[\sum_{k\in U} I_k\right] = \sum_{k\in U}\sum_{l\in U} \operatorname{cov}(I_k, I_l) = \\ &= \sum_{k\in U} \mathcal{D}(I_k) + \sum_{k\in U}\sum_{l\neq k} \operatorname{cov}(I_k, I_l) = \\ &= \sum_{k\in U} \pi_k(1-\pi_k) + \sum_{k\in U}\sum_{l\neq k} (\pi_{kl} - \pi_k\pi_l) = \\ &= \sum_{k\in U} \pi_k - \left[\sum_{k\in U} \pi_k\right]^2 + \sum_{k\in U}\sum_{l\neq k} \pi_{kl}. \end{aligned} \tag{1.3}$$

Приклад 1.10. У випадку застосування відбору Бернуллі розмір вибірки – це біноміально розподілена випадкова величина з параметрами N та π. Тобто,

$$E_{\text{ВБ}}(n_{\mathcal{S}}) = N\pi, \quad \mathcal{D}_{\text{ВБ}}(n_{\mathcal{S}}) = N\pi(1-\pi).$$

Ці формули можна отримати з (1.2) та (1.3) враховуючи, що $\pi_k = \pi$ для всіх k та $\pi_{kl} = \pi^2$ для $k \neq l$. Зокрема,

$$\mathcal{D}_{\text{ВБ}}(n_{\mathcal{S}}) = N\pi - (N\pi)^2 + N(N-1)\pi^2 = N\pi(1-\pi).$$

◊

На практиці вибіркові дизайни зі змінним розміром вибірки використовуються рідко, оскільки зазвичай у цьому випадку оцінки мають більшу дисперсію. А також, що є більш важливим, при плануванні обстеження бажано знати наперед, яку кількість елементів буде обстежено.

Означення 1.4. *Вибірковим дизайном із фіксованим розміром вибірки* називається такий дизайн, що при $p(s) > 0$ вибірка s складається з фіксованої кількості елементів, наприклад, n. Але це не означає, що всі вибірки розміру n можуть бути обраними. Тобто, деякі вибірки розміру n можуть мати нульову ймовірність бути обраними.

Для *вибіркових дизайнів з фіксованим розміром вибірки* ймовірності включення першого та другого порядків мають властивості, сформульовані в твердженні 1.2.

***Твердження* 1.2.** *Якщо вибірковий дизайн $p(\cdot)$ має фіксований розмір n, то*

$$\begin{aligned} \sum_{k\in U} \pi_k &= n; \\ \sum_{k\in U}\sum_{l\neq k} \pi_{kl} &= n(n-1); \\ \sum_{l\in U, l\neq k} \pi_{kl} &= (n-1)\pi_k \quad \forall k. \end{aligned}$$

Доведення. Із того, що вибірковий дизайн $p(\cdot)$ є дизайном із фіксованим розміром вибірки, випливає, що $E(n_s) = n$ та $\mathcal{D}(n_s) = 0$. Із (1.2) та (1.3) отримуємо перші два співвідношення та

$$\begin{aligned} \forall k \in U \quad \sum_{l\in U, l\neq k} \pi_{kl} &= \sum_{l\in U, l\neq k} E[I_k I_l] = E\left[I_k\left(\sum_{l\in U} I_l - I_k\right)\right] = \\ &= nE[I_k] - E[I_k^2] = (n-1)\pi_k. \end{aligned}$$

□

1.7. Оцінка Горвіца–Томпсона та її властивості

Розглянемо та дослідимо оцінку для сумарного значення T досліджуваної характеристики y генеральної сукупності, яка називається *оцінкою Горвіца–Томпсона* або просто *π-оцінкою*:

$$\widehat{t}_\pi = \sum_{k\in s} \frac{y_k}{\pi_k}. \tag{1.4}$$

***Твердження* 1.3.** *π-оцінка $\widehat{t}_\pi$ є незміщеною оцінкою сумарного значення T з дисперсією*

$$\mathcal{D}(\widehat{t}_\pi) = \sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum_{k\in U}\sum_{l\in U}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) y_k y_l. \tag{1.5}$$

Якщо $\pi_{kl} > 0$ для всіх $k, l \in U$, то незміщеною оцінкою дисперсії буде статистика

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum_{k\in s}\sum_{l\in s}\frac{1}{\pi_{kl}}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) y_k y_l. \tag{1.6}$$

Доведення. Скористаймося записом

$$\widehat{t}_\pi = \sum_{k\in U} I_k \frac{y_k}{\pi_k}.$$

Із того, що $E(I_k) = \pi_k$ та $\pi_k > 0$ для всіх $k \in U$, випливає, що $\widehat{t}_\pi$ є незміщеною оцінкою параметра T.

Оскільки $\pi_{kk} = \pi_k$ та

$$\sum_{k\in U}\sum_{l\in U} a_{kl} = \sum_{k\in U} a_{kk} + \sum_{k\in U}\sum_{\substack{l\in U:\\ l\neq k}} a_{kl},$$

то

$$\mathcal{D}\left(\widehat{t}_\pi\right) = \mathcal{D}\left(\sum_{k\in U} I_k \frac{y_k}{\pi_k}\right) =$$

$$= \sum_{k\in U} \mathcal{D}\left(I_k \frac{y_k}{\pi_k}\right) + \sum_{k\in U} \sum_{\substack{l\in U:\\ l\neq k}} \operatorname{cov}\left(I_k \frac{y_k}{\pi_k}, I_l \frac{y_l}{\pi_l}\right) =$$

$$= \sum_{k\in U} \left(\frac{y_k}{\pi_k}\right)^2 \pi_k (1-\pi_k) + \sum_{k\in U} \sum_{\substack{l\in U:\\ l\neq k}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_{kl} - \pi_k\pi_l) =$$

$$= \sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Залишилось довести незміщеність оцінки дисперсії, яку можна записати як

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = \sum_{k\in U}\sum_{l\in U} I_k I_l \left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\right)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Оскільки для всіх $k, l \in U$, для яких $\pi_{kl} > 0$, виконується рівність

$$E\left[I_k I_l \left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\right)\right] = \pi_{kl}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} = \pi_{kl} - \pi_k\pi_l,$$

то

$$E\widehat{\mathcal{D}}(\widehat{t}_\pi) = \mathcal{D}(\widehat{t}_\pi),$$

що й треба було довести. □

Для вибіркових дизайнів з *фіксованим розміром вибірки* дисперсію можна записати дещо по-іншому. При цьому оцінка для дисперсії буде мати також інший вигляд.

***Твердження* 1.4. Форма Єйтса–Гранді–Сена.** *Якщо $p(\cdot)$ – вибірковий дизайн з фіксованим розміром вибірки, то дисперсію оцінки Горвіца–Томпсона можна подати у вигляді*

$$\mathcal{D}(\widehat{t}_\pi) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2.$$

Якщо для всіх $k, l \in U$, $\pi_{kl} > 0$, то незміщеною оцінкою дисперсії буде статистика

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2.$$

Доведення. Перевіримо спочатку еквівалентність виразів для $\mathcal{D}(\widehat{t}_\pi)$ за умови фіксованого розміру вибірки n. Справді

$$\begin{aligned}
\mathcal{D}(\widehat{t}_\pi) &= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 = \\
&= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\left[\left(\frac{y_k}{\pi_k}\right)^2 - 2\frac{y_k y_l}{\pi_k\pi_l} + \left(\frac{y_l}{\pi_l}\right)^2\right] = \\
&= \sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\frac{y_k y_l}{\pi_k\pi_l} - \sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\left(\frac{y_k}{\pi_k}\right)^2
\end{aligned}$$

та

$$\sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\left(\frac{y_k}{\pi_k}\right)^2 = \sum_{k\in U}\left(\left(\frac{y_k}{\pi_k}\right)^2 \sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\right).$$

І оскільки для будь-якого фіксованого $k \in U$: $\sum\limits_{l\in U} (\pi_{kl} - \pi_k\pi_l) = \sum\limits_{l\in U} \pi_{kl} - \pi_k \sum\limits_{l\in U} \pi_l = (n-1)\pi_k + \pi_k - \pi_k n = 0$, то ці дві форми еквівалентні для дизайну з фіксованим розміром вибірки.

Незміщенність оцінки дисперсії доводиться так само, як і в доведенні твердження 1.3. □

Зауваження 1.2. Далі ми будемо дуже часто зустрічатись у викладках з виразом $\pi_{kl} - \pi_k\pi_l$, тому для економії часу і місця для нього варто ввести спеціальне позначення $\Delta_{kl} := \pi_{kl} - \pi_k\pi_l$.

Зауваження 1.3. Для дизайну з фіксованим розміром вибірки значення дисперсій $\mathcal{D}(\widehat{t}_\pi)$ в обох формах завжди будуть однакові, а значення оцінок $\widehat{\mathcal{D}}(\widehat{t}_\pi)$ можуть відрізнятись залежно від того, яка форма використовується. Але обидві ці оцінки є незміщеними.

Зауваження 1.4. Залежно від набору значень y_k оцінки дисперсій як у формі Єйтса–Гранді–Сена, так і у формі Горвіца–Томпсона можуть набувати від'ємних значень. Достатньою умовою невід'ємності цих оцінок є *умова Єйтса–Гранді–Сена*:

$$\Delta_{kl} \le 0 \quad \text{для всіх } k \ne l \in U.$$

Сумарне значення T та середнє значення $\overline{Y}$ досліджуваної характеристики y генеральної сукупності пов'язані між собою співвідношенням

$$T = \sum_{k\in U} y_k = N\left(\frac{1}{N}\sum_{k\in U} y_k\right) = N\overline{Y}.$$

Звідси можна легко отримати оцінку Горвіца–Томпсона для середнього значення $\overline{Y}$ та відповідні результати для дисперсії цієї оцінки.

Наслідок 1.1. *π-оцінка $\widehat{\overline{y}}_\pi = \frac{1}{N}\sum\limits_{k\in s}\frac{y_k}{\pi_k}$ є незміщеною оцінкою середнього $\overline{Y}$ з дисперсією*

$$\mathcal{D}(\widehat{\overline{y}}_\pi) = \mathcal{D}\left(\frac{1}{N}\widehat{t}_\pi\right) = \frac{1}{N^2}\mathcal{D}(\widehat{t}_\pi) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Якщо $\pi_{kl} > 0$ для всіх $k, l \in U$, то незміщеною оцінкою дисперсії буде статистика

$$\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi) = \widehat{\mathcal{D}}\left(\frac{1}{N}\widehat{t}_\pi\right) = \frac{1}{N^2}\widehat{\mathcal{D}}(\widehat{t}_\pi) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

1.8. Вправи та питання для самоконтролю

1.1. У чому полягає суттєва різниця між методами відбору за допомогою жеребкування та послідовного відбору зі списку?

1.2. Що є стратегією відбору? Чи визначається вона однозначно?

1.3. Чи може оцінка дисперсії π-оцінки набувати від'ємних значень? А у формі Єйтса–Гранді–Сена?

1.4. Проводиться обстеження деякого регіону України з метою визначення середнього доходу домогосподарств. Якщо для проведення відбору немає списку всіх домогосподарств, але доступна інформація з реєстру осіб, що проживають у даному регіоні, то можна застосувати такий метод відбору: простим випадковим відбором з N осіб відбираються n осіб, а потім визначають домогосподарства, до яких належать відібрані особи.

Підрахувати ймовірність включення у вибірку домогосподарства, що складається з M осіб ($M < n$). Отримати наближений вираз для цієї ймовірності при $M = 1, 2, 3$, припускаючи, що n і N є досить великими та $\frac{n}{N} = f > 0$.

1.5. Розглянемо генеральну сукупність $U = \{1, 2, 3\}$ з вибірковим дизайном $p(\cdot)$, при якому

$$p(\{1,2\}) = \frac{1}{2}; \quad p(\{1,3\}) = \frac{1}{4}, \quad p(\{2,3\}) = \frac{1}{4}.$$

Знайти ймовірності включення першого порядку та коваріаційну матрицю індикаторів включення $\Delta = \{\Delta_{kl} = \operatorname{cov}(I_k, I_l)\}_{k,l\in U}$.

1.6. Нехай коваріаційна матриця індикаторів включення при деякому дизайні $p(\cdot)$ має вигляд

$$\Delta = \frac{6}{25} \times \begin{pmatrix} 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{pmatrix}.$$

1) Чи є $p(\cdot)$ дизайном з фіксованим розміром вибірки?
2) Чи задовольняє цей дизайн умову Єйтса–Гранді–Сена?
3) Обчислити ймовірності включення першого порядку для дизайну $p(\cdot)$, якщо $\pi_1 = \pi_2 = \pi_3 > \pi_4 = \pi_5$.
4) Записати матрицю ймовірностей включення другого порядку. *Вказівка*: скористайтеся тим, що $\pi_{kl} = \Delta_{kl} - \pi_k\pi_l$ для всіх $k, l \in U$.
5) Обчислити ймовірності отримання всіх можливих вибірок.

1.7. Розглянемо генеральну сукупність, що складається з п'яти елементів: $U = \{1, 2, 3, 4, 5\}$. При вибірковому дизайні $p(\cdot)$ можливими вибірками є: $s_1 = \{1, 2\}$; $s_2 = \{1, 2, 3\}$; $s_3 = \{2, 3, 4\}$ та

$s_4 = \{1, 2, 3, 4, 5\}$. При цьому $p(s_1) = 0,1$; $p(s_2) = 0,2$; $p(s_3) = 0,3$ і $p(s_4) = 0,4$.

1) Обчислити всі ймовірності включення першого та другого порядку.

2) Знайти En_s та $\mathcal{D}n_s$, використовуючи означення математичного сподівання та дисперсії дискретної випадкової величини.

3) Знайти En_s та $\mathcal{D}n_s$, використовуючи формули (1.2) та (1.3).

1.8. Генеральна сукупність, що складається з 1600 осіб, поділена на 800 груп (домогосподарств) так, що є N_i груп розміру i, $i = 1, 2, 3, 4$.

i	1	2	3	4
N_i	250	350	150	50

Вибірка осіб утворюється простим випадковим відбором без повернення. Відбираються 300 домогосподарств із 800 і обстежуються всі особи, що належать до вибраних домогосподарств. Нехай n_s – це загальна кількість осіб у вибірці. Обчислити En_s та $\mathcal{D}n_s$.

1.9. Нехай для генеральної сукупності та вибіркового дизайну з вправи 1.8 характеристика y має такі значення: $y_1 = 10$, $y_2 = 5$, $y_3 = 7$, $y_4 = 3$ та $y_5 = 1$. Знайти:

1) математичне сподівання та дисперсію оцінки Горвіца–Томпсона для сумарного T;

2) коефіцієнт варіації оцінки $\widehat{t}_\pi$;

3) оцінку дисперсії $\widehat{\mathcal{D}}(\widehat{t}_\pi)$ для кожної з чотирьох можливих вибірок;

4) математичне сподівання оцінки дисперсії $\widehat{\mathcal{D}}(\widehat{t}_\pi)$.

Розділ 2

Простий випадковий відбір без повернення

2.1. Основні властивості

У цьому розділі ми розглянемо основні характеристики простого випадкового відбору без повернення (ПВВбП). Цей метод відбору вже дещо знайомий вам з прикладів попереднього розділу.

У випадку застосування простого випадкового відбору *без повернення* індикатори включення елементів у вибірку $I_1, I_2, ..., I_k$ не є незалежними. Вибірковий дизайн ПВВбП має вигляд (див. приклад 1.3):

$$p(s) = \begin{cases} \frac{1}{C_N^n}, & \text{якщо розмір вибірки дорівнює } n; \\ 0, & \text{у протилежному випадку.} \end{cases}$$

Ймовірності включення однакові для всіх елементів генеральної сукупності і мають вигляд:

$$\begin{aligned} \pi_k &= \frac{n}{N}, \quad \forall k = \overline{1, N}; \\ \pi_{kl} &= \frac{n(n-1)}{N(N-1)}, \quad k \neq l; \\ \pi_{kk} &= \pi_k = \frac{n}{N}. \end{aligned}$$

Для того, щоб отримати вибірковий дизайн ПВВбП, можна використовувати різні ймовірнісні схеми відбору. Однією з найбільш поширених та простих є схема відбору на основі жеребкування. Але у випадку, коли інформація про елементи генеральної сукупності міститься в деякому файлі та відбір бажано алгоритмізувати, більш зручними є схеми послідовного відбору зі списку. Наведемо декілька таких схем відбору.

Схема відбору 1. Нехай $\varepsilon_1, \varepsilon_2, ..., \varepsilon_N$ – незалежні реалізації рівномірно розподіленої на $[0, 1]$ випадкової величини

($\varepsilon \sim Unif[0,1]$). Кожному елементу k генеральної сукупності ставимо у відповідність випадкове значення ε_k. Далі перебираємо по черзі елементи генеральної сукупності та вирішуємо включати елемент чи ні з імовірністю, що дорівнює відношенню кількості елементів, що залишилось включити у вибірку, до кількості елементів, що залишилось переглянути.

Крок 1. Якщо $\varepsilon_1 < n/N$, то елемент 1 відбирається; якщо $\varepsilon_1 \geq n/N$, то ні.

Крок k. Якщо $\varepsilon_k < \frac{n-n_k}{N-k+1}$, то k-й елемент відбирається; якщо $\varepsilon_k \geq \frac{n-n_k}{N-k+1}$, то ні. Тут n_k – це кількість елементів відібраних у вибірку з перших $k-1$ елементів зі списку.

Процедура закінчується, коли $n_k = n$.

У результаті такої схеми відбору отримуємо вибірковий дизайн ПВВбП.

У випадку, коли розмір генеральної сукупності невідомий, а це часто буває, для отримання дизайну ПВВбП можна використати іншу схему відбору.

Схема відбору 2. Спочатку до вибірки потрапляють перші n елементів генеральної сукупності. А потім розглядається кожен наступний елемент по черзі, починаючи з $n+1$.

Крок 1. Елемент $n+1$ потрапляє у вибірку з імовірністю $\frac{n}{n+1}$. Якщо елемент $n+1$ потрапив таким чином у вибірку, тоді з неї викидається один елемент, вибраний випадковим чином з однаковою для всіх імовірністю. У результаті у вибірці знову залишається n елементів.

Крок k. Для всіх інших елементів k, $n+1 < k \leq N$, ймовірність потрапити у вибірку буде дорівнювати $\frac{n}{k}$, та якщо елемент k вибирається, то з вибірки викидається один елемент з однаковою для всіх імовірністю.

Така схема відбору може бути реалізована навіть без відомого наперед розміру генеральної сукупності N. Формування списку

елементів сукупності може бути одночасним із процедурою відбору.

Схема відбору 3. Генерується N незалежних, рівномірно розподілених на [0,1] випадкових величин $\varepsilon_1, \varepsilon_2, ..., \varepsilon_N$, де кожному елементу $k \in U$ відповідає значення ε_k. Потім ці значення впорядковуються, наприклад, по зростанню і діляться на вибірки розміру n. Кожна така вибірка не перетинається з іншою та продукує вибірковий дизайн ПВВбП.

Основна перевага цієї схеми відбору полягає у можливості одночасно вибрати декілька простих випадкових вибірок, що не перетинаються. Це дуже важливо у випадку, коли проводиться серія обстежень однієї і тієї самої генеральної сукупності протягом короткого проміжку часу. Таким чином можливо зменшити рівень невідповідей за рахунок зменшення навантаження на респондентів. Такі вибірки називаються *від'ємно-координованими*.

2.2. Оцінка Горвіца–Томпсона при простому випадковому відборі без повернення

Застосуємо метод Горвіца–Томпсона для оцінювання сумарного значення досліджуваної характеристики генеральної сукупності у випадку простого випадкового відбору без повернення.

Згідно з формулою (1.4) при ПВВбП π-оцінка Горвіца–Томпсона матиме такий вигляд:

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{N}{n} y_k = N \left(\frac{1}{n} \sum_{k \in s} y_k \right) = N \overline{y},$$

де $\overline{y} = \frac{1}{n} \sum_{k \in s} y_k$ – вибіркове середнє.

Оскільки вибірковий дизайн ПВВбП має фіксований розмір вибірки, то можна скористатися формулою Єйтса–Гранді–Сена для знаходження дисперсії та оцінки дисперсії π-оцінки Горвіца–

Томпсона. Але спочатку для всіх $k \neq l$ обчислимо

$$\begin{aligned}\Delta_{kl} &= \pi_{kl} - \pi_k\pi_l = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = \\ &= \frac{n}{N}\left(\frac{(n-1)N - n(N-1)}{N(N-1)}\right) = \frac{n}{N}\frac{n-N}{N(N-1)} = \\ &= \frac{n}{N}\left(\frac{n}{N} - 1\right)\frac{1}{N-1} = -\frac{f(1-f)}{N-1},\end{aligned}$$

де через $f := \frac{n}{N}$ будемо позначати *частку відбору*.

Інтуїтивно зрозуміло, що дисперсія оцінки залежить від того, наскільки однорідними є елементи генеральної сукупності. Параметром, що відображає цю властивість, є дисперсія генеральної сукупності

$$S_y^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \overline{Y})^2 = \frac{1}{N-1}\sum_{k\in U} y_k^2 - \frac{1}{N(N-1)}\left(\sum_{k\in U} y_k\right)^2.$$

Нижній індекс у позначенні дисперсії S_y^2 вказує, що це дисперсія саме характеристики y, оскільки при обстеженні однієї і тієї самої генеральної сукупності одночасно можуть досліджуватись декілька характеристик: $x, y, z, \ldots$. Якщо досліджується тільки одна змінна, то цей індекс можна не писати.

Отже, виведемо формулу для обчислення дисперсії π-оцінки $\widehat{t}_\pi$ при ПВВбП:

$$\begin{aligned}\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi) &= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 = \\ &= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\left(-\frac{f(1-f)}{N-1}\right)\left(\frac{N}{n}\right)^2(y_k - y_l)^2 = \\ &= \frac{1-f}{2(N-1)f}\left[2N\sum_{k\in U} y_k^2 - 2\sum_{k\in U} y_k\sum_{l\in U} y_l\right] = \\ &= \frac{1-f}{f}N\left[\frac{1}{N-1}\sum_{k\in U} y_k^2 - \frac{1}{(N-1)N}\left(\sum_{k\in U} y_k\right)^2\right] = \end{aligned}$$

$$= \frac{N}{n}(1-f)NS^2 = N^2(1-f)\frac{S^2}{n}.$$

Аналогічно отримуємо вираз для обчислення незміщеної оцінки цієї дисперсії

$$\widehat{\mathcal{D}}_{\text{ПВВбП}}(\widehat{t}_\pi) = N^2\frac{1-f}{n}\widehat{S}^2,$$

де $\widehat{S}^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \overline{y})^2$ – вибіркова дисперсія (дисперсія елементів, що потрапили у вибірку) змінної y. Зауважимо, що $E\widehat{S}^2 = S^2$.

Отже, тепер сформулюємо твердження.

***Твердження* 2.1.** *При простому випадковому відборі без повернення оцінки Горвіца–Томпсона сумарного та середнього значень набувають вигляду:*

$$\begin{aligned}\widehat{t}_\pi &= \frac{N}{n}\sum_{k\in s} y_k = \frac{N}{n}t, \text{ де } t = \sum_{k\in s} y_k - \text{ вибіркове сумарне;} \\ \widehat{\overline{y}}_\pi &= \frac{1}{N}\widehat{t}_\pi = \frac{t}{n} = \frac{1}{n}\sum_{k\in s} y_k = \overline{y}, \text{ де } \overline{y} - \text{ вибіркове середнє.}\end{aligned}$$

Запишемо дисперсії цих оцінок:

$$\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi) = N^2\frac{(1-f)}{n}S^2, \quad \mathcal{D}_{\text{ПВВбП}}(\widehat{\overline{y}}_\pi) = \frac{(1-f)}{n}S^2. \quad (2.1)$$

Незміщені оцінки дисперсій обчислюються так:

$$\widehat{\mathcal{D}}_{\text{ПВВбП}}(\widehat{t}_\pi) = N^2\frac{(1-f)}{n}\widehat{S}^2, \quad \widehat{\mathcal{D}}_{\text{ПВВбП}}(\widehat{\overline{y}}_\pi) = \frac{(1-f)}{n}\widehat{S}^2, \quad (2.2)$$

де S^2 – дисперсія елементів у генеральній сукупності, $\widehat{S}^2$ – дисперсія елементів у вибірці, $f = \frac{n}{N}$ – частка відбору.

Зауваження 2.1. Для простого випадкового відбору та для стратифікованого відбору формули для дисперсій у формі Горвіца–Томпсона та Єйтса–Гранді–Сена є ідентичними.

Продемонструємо на прикладі, як працює наведена теорія.

Приклад 2.1. Обстежується генеральна сукупність, що складається з шести осіб ($N = 6$), яких назвемо A, B, C, D, E, F (табл. 1). Потрібно оцінити середнє значення характеристики y на основі вибірки, що складається з двох осіб ($n = 2$). Характеристикою y може бути, наприклад, зріст, вага, дохід за місяць, лояльність до деякого кандидата на виборах і т. д.

Припустимо також, що значення характеристики y відомі для всіх осіб генеральної сукупності. Це, звичайно, припущення, яке ніколи не виконується на практиці. Інакше, навіщо тоді проводити вибіркове обстеження, якщо відома вся інформація про генеральну сукупність? Але це припущення потрібне для того, щоб проілюструвати механізм вибіркових обстежень.

Таблиця 1. **Елементи генеральної сукупності з відповідними значеннями характеристики y**

Особа	A	B	C	D	E	F
k	1	2	3	4	5	6
y_k	2	6	8	10	10	12

Якщо застосувати схему відбору, яка полягає у послідовному рівноймовірнісному відборі двох із шести елементів генеральної сукупності без повернення вже вибраних елементів у сукупність, то отримаємо вибірковий дизайн, що відповідає простому випадковому відбору без повернення:

$$p(s) = \begin{cases} \frac{1}{15}, & \text{якщо розмір вибірки } n = 2; \\ 0, & \text{в інших випадках.} \end{cases}$$

При цьому елементи, що потрапляють у вибірку, будуть різними. Всього таких вибірок буде $C_6^2 = 15$. У табл. 2 занесені всі такі вибірки та обчислені відповідні вибіркові середні, які є оцінками для середнього значення $\overline{Y} = 8$ характеристики y генеральної сукупності.

Таблиця 2. **15 можливих вибірок без повернення розміру 2 із сукупності розміру 6 з вибірковими середніми $\overline{y}$**

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Елемент	A B	A C	A D	A E	A F	B C	B D	B E	B F	C D	C E	C F	D E	D F	E F
Значення	2 6	2 8	2 10	2 10	2 12	6 8	6 10	6 10	6 12	8 10	8 10	8 12	10 10	10 12	10 12
$\overline{y}$	4	5	6	6	7	7	8	8	9	9	9	10	10	11	11

Ймовірності включення першого та другого порядків при цьому будуть однаковими для всіх елементів

$$\forall k = \overline{1,6}: \quad \pi_k = \frac{2}{6} = \frac{1}{3}; \quad \forall k \neq l: \quad \pi_{kl} = \frac{1}{15},$$

тобто, всі елементи мають однаковий шанс потрапити у вибірку.

Якщо уважно подивитись на всі вибірки, то можна помітити, що деякі з них погано відображають властивості генеральної сукупності. Так, наприклад, вибірка 1, що включає значення 2 та 6. Середнє по цій вибірці $\overline{y} = 4$ досить сильно відрізняється від реального значення середнього по генеральній сукупності $\overline{Y} = 8$. Але ця вибірка має такий самий, як і всі інші, шанс бути обраною. Це означає, що випадковий відбір не гарантує на 100 %, що оцінка по кожній вибірці буде мати значення близьке до реального значення параметра. Ми повинні змиритися з тим фактом, що отримана нами вибірка може бути досить «поганою». Питання в тому, як багато таких вибірок при певних видах відбору і який шанс їх отримати.

Як видно з табл. 2 значення оцінки середнього $\widehat{\overline{y}}_\pi$ – вибіркового середнього $\overline{y}$, змінюються залежно від того, які елементи потрапляють у вибірку. Тобто, $\overline{y}$ – це дискретна випадкова величина. Запишемо її розподіл (табл. 3).

Таблиця 3. **Вибірковий розподіл $\overline{y}$**

$\overline{y}$	4	5	6	7	8	9	10	11
P	1/15	1/15	2/15	2/15	2/15	3/15	2/15	2/15

Ймовірності, що відповідають можливим значенням оцінки обчислюються за формулою (1.1).

Застосувавши класичне визначення математичного сподівання дискретної випадкової величини, можна впевнитись, що оцінка $\widehat{\overline{y}}_\pi$ дійсно є незміщеною, тобто її математичне сподівання співпадає з реальним значенням параметра, що оцінюється. Дисперсію цієї оцінки можна порахувати в цьому випадку двома способами: за допомогою визначення дисперсії дискретної випадкової величини та за формулою (2.1). Отримаємо 64/15.

Але на практиці ми ніколи не будемо знати значення характеристики y для всіх елементів генеральної сукупності, а отже і не зможемо побудувати вибірковий розподіл оцінюючої статистики. Тому для визначення дисперсії оцінки ми не зможемо скористатись ні означенням, ні формулою (2.1). Єдине, що ми можемо зробити – це оцінити дисперсію за формулою (2.2). ◊

2.3. Оцінювання параметрів підсукупностей при ПВВбП

Часто метою обстеження є виявлення та дослідження елементів, що належать до деякої підсукупності. Наприклад, підсукупністю, що потребує вивчення, може бути безробітне населення України – кількість безробітних та відносна частка таких людей у населенні України. У цьому випадку генеральну сукупність ділять на підсукупності і вивчають кожну з них.

Позначимо підсукупність $U_d \subset U$ (d від англійського терміну `domain` – підсукупність). Нехай N_d – кількість елементів, що належать даній підсукупності, тоді $P_d = \frac{N_d}{N}$ – частка елементів у генеральній сукупності U, що належать підсукупності U_d. Припустимо, що розмір генеральної сукупності N відомий, а розмір підсукупності N_d – ні, що часто зустрічається на практиці.

2.3.1. Оцінювання абсолютного та відносного розмірів підсукупності

Для оцінки параметрів N_d та P_d можемо скористатись підходами, подібними до оцінювання сумарного та середнього значень генеральної сукупності.

Для цього введемо змінну z_{dk} – індикатор належності елемента k до сукупності U_d:

$$z_{dk} = \begin{cases} 1, k \in U_d, \\ 0, k \notin U_d, \end{cases} \qquad k = 1, ..., N.$$

Тоді

$$N_d = \sum_{k \in U} z_{dk} \quad \text{та} \quad P_d = \frac{N_d}{N} = \frac{1}{N} \sum_{k \in U} z_{dk} = \overline{z}_d.$$

Тобто, N_d – це сумарне, а P_d – середнє значення нової змінної z_{dk}. Отже, для цих параметрів справедливі всі результати, отримані нами раніше. Але у зв'язку з особливим виглядом нової змінної – вона є дихотомічною, варто записати ці результати, використовуючи трохи іншу термінологію.

Нехай $Q_d = 1 - P_d$, $n_d = \sum_{k \in s} z_{dk}$ – кількість елементів у вибірці, що належать підсукупності U_d, $p_d = \frac{n_d}{n}$ – частка елементів у вибірці, що належать підсукупності U_d, $q_d = 1 - p_d$. Тоді

$$\begin{aligned} S^2_{z_d} &= \frac{1}{N-1} \sum_{k \in U} z^2_{kd} - \frac{1}{N(N-1)} \left(\sum_{k \in U} z_{dk} \right)^2 = \\ &= \frac{1}{N-1} N_d - \frac{1}{N(N-1)} N_d^2 = \\ &= \frac{N_d}{N-1} \left(1 - \frac{N_d}{N} \right) = \frac{NP_d}{N-1} Q_d = \frac{N}{N-1} P_d Q_d \end{aligned}$$

і, аналогічно

$$\widehat{S}^2_{z_d} = \frac{n}{n-1} p_d q_d.$$

***Твердження* 2.2.** *При використанні простого випадкового відбору без повернення оцінка Горвіца–Томпсона для розміру N_d підсукупності U_d та її дисперсія набувають вигляду:*

$$\begin{aligned}\widehat{N}_d &= Np_d;\\ \mathcal{D}_{\text{ПВВбП}}\left(\widehat{N}_d\right) &= N^2\left(\frac{1}{n}-\frac{1}{N}\right)S^2_{z_d} = N^2\frac{N-n}{N-1}\frac{P_dQ_d}{n};\\ \widehat{\mathcal{D}}_{\text{ПВВбП}}\left(\widehat{N}_d\right) &= N^2(1-f)\frac{p_dq_d}{n-1}.\end{aligned}$$

Аналогічно отримується оцінка для P_d.

***Твердження* 2.3.** *При використанні простого випадкового відбору без повернення оцінка Горвіца–Томпсона частки P_d елементів, що належать заданій підсукупності U_d, має вигляд:*

$$\begin{aligned}\widehat{P}_d &= p_d;\\ \mathcal{D}_{\text{ПВВбП}}\left(\widehat{P}_d\right) &= \frac{N-n}{N-1}\frac{P_dQ_d}{n};\\ \widehat{\mathcal{D}}_{\text{ПВВбП}}\left(\widehat{P}_d\right) &= (1-f)\frac{p_dq_d}{n-1}.\end{aligned}$$

Зауваження 2.2. При визначенні розміру вибірки можна використати той факт, що свого максимального значення 1/4 вираз $P_dQ_d = P_d(1-P_d)$ набуває, коли $P_d = 1/2$.

2.3.2. Оцінювання сумарного та середнього підсукупності, коли її розмір N_d невідомий

Розглянемо ще один тип задач, коли потрібно оцінити сумарне значення $T_d = \sum_{k\in U_d} y_k$ та середнє значення $\overline{Y}_d = \frac{1}{N_d}\sum_{k\in U_d} y_k$ характеристики y у підсукупності U_d.

Введемо нову змінну

$$y_{dk} = \begin{cases} y_k, & k \in U_d;\\ 0, & k \notin U_d.\end{cases}$$

Тоді T_d – це сумарне значення по всій генеральній сукупності U нових змінних y_{dk}:

$$T_d = \sum_{k\in U_d} y_k = \sum_{k\in U} y_{dk},$$

для якого оцінка Горвіца–Томпсона при простому випадковому відборі набуває вигляду

$$\widehat{t}_{d\pi} = \sum_{k\in s}\frac{y_{dk}}{\pi_k} = \frac{N}{n}\sum_{k\in s} y_{dk} = \frac{N}{n}\sum_{k\in s_d} y_k, \quad \text{де} \quad s_d = s\cap U_d.$$

Вирази для $\mathcal{D}_{\text{ПВВбП}}\left(\widehat{t}_{d\pi}\right)$ та $\widehat{\mathcal{D}}_{\text{ПВВбП}}\left(\widehat{t}_{d\pi}\right)$ отримуються з відповідних формул для оцінки Горвіца–Томпсона.

У випадках, коли розмір підсукупності N_d відомий, зазвичай для T_d використовують оцінку, альтернативну до π-оцінки:

$$\widehat{t}_{d,\text{ альт}} = N_d\left(\frac{1}{n_d}\sum_{k\in s_d} y_k\right) = N_d\overline{y}_d,$$

де n_d – розмір підвибірки s_d. Але в цьому випадку n_d – це випадкова величина і, щоб отримати вираз для дисперсії $\widehat{t}_{d,\text{ альт}}$ (що зазвичай менша за $\widehat{t}_{d\pi}$), потрібно застосувати інший метод, який буде розглянуто в пункті 9.3.

Для оцінювання $\overline{Y}_d = \frac{T_d}{N_d}$ скористаємося π-оцінкою

$$\widehat{\overline{y}}_{d,\pi} = \frac{1}{N_d}\left(\frac{N}{n}\sum_{k\in s_d} y_k\right).$$

Якщо величина N_d – невідома, то використовувати $\widehat{\overline{y}}_{d,\pi}$ не можна. Кращою оцінкою для $\overline{Y}_d$, що може бути використана незалежно від того, відоме N_d чи ні, є:

$$\widehat{\overline{y}}_{d,\text{ альт}} = \frac{1}{N_d}\widehat{t}_{d,\text{ альт}} = \frac{1}{n_d}\sum_{k\in s_d} y_k = \overline{y}_d.$$

2.4. Побудова довірчих інтервалів

Довірчим інтервалом для параметра θ називається інтервал, що визначається парою статистик $\widehat{\theta}_{\text{Н}}(s)$ та $\widehat{\theta}_{\text{В}}(s)$, таких, що $\widehat{\theta}_{\text{Н}}(s) \le \widehat{\theta}_{\text{В}}(s)$ для будь-якої вибірки s та

$$P\{\theta \in [\widehat{\theta}_{\text{Н}}(s), \widehat{\theta}_{\text{В}}(s)]\} \le 1-\alpha.$$

Довірчі інтервали позначимо $DI(s)$. Величину $1-\alpha$ назвемо *рівнем довіри*, а $\widehat{\theta}_{\text{В}}(s) - \widehat{\theta}_{\text{Н}}(s)$ – *точністю* (шириною) довірчого інтервалу.

Довірчі інтервали в контексті теорії вибіркових обстежень інтерпретуються таким чином. Припустимо, що відомі всі N значень змінної y, тоді значення θ буде відоме і ми можемо побудувати довірчі інтервали $DI(s)$ для кожної можливої вибірки s ($p(s) > 0$). Тоді для кожної такої вибірки можемо визначити, чи належить реальне значення параметра θ відповідному довірчому інтервалу, чи ні, та підрахувати

$$P[DI(s) \ni \theta] = 1 - \alpha,$$

де $\alpha = \sum\limits_{s \in S^*} p(s)$ – це сумарна ймовірність по множині S^* тих вибірок, довірчі інтервали яких не містять параметр θ.

Приклад 2.2. Розглянемо генеральну сукупність, що складається з трьох елементів, для яких відомі значення характеристики y: $y_1 = 3$, $y_2 = 7$, $y_3 = 5$. Характеристикою, що нас цікавить, є сумарне значення $T = \sum\limits_{k \in U} y_k = 15$.

Простим випадковим відбором без повернення можемо отримати 3 різні вибірки розміру 2:

$$s_1 = \{1,2\}, \quad s_2 = \{1,3\}, \quad s_3 = \{2,3\}.$$

Нехай довірчий інтервал для сумарного будується таким чином:

$$DI(s) = \left[\widehat{\theta}_{\text{Н}}(s) = \widehat{t}_\pi - \sqrt{\widehat{\mathcal{D}}(\widehat{t}_\pi)};\ \widehat{\theta}_{\text{В}}(s) = \widehat{t}_\pi + \sqrt{\widehat{\mathcal{D}}(\widehat{t}_\pi)}\right],$$

де $\widehat{t}_\pi$ – оцінка Горвіца–Томпсона при ПВВбП, $\widehat{\mathcal{D}}(\widehat{t}_\pi)$ – оцінка її дисперсії.

Для кожної вибірки побудуємо довірчі інтервали

$$\begin{aligned} DI(s_1) &= 15 \pm \sqrt{12} = [11,54; 18,46]; \\ DI(s_2) &= 12 \pm \sqrt{3} = [10,27; 13,73]; \\ DI(s_3) &= 18 \pm \sqrt{3} = [16,54; 18,46]. \end{aligned}$$

Реальне значення $T = 15$ потрапляє лише в один довірчий інтервал. Отже, рівень довіри для такого довірчого інтервалу $1 - \alpha = 1/3 = 0,33$.

Зрозуміло, що такий рівень довіри замалий. У цьому випадку можливо або збільшувати розмір вибірки, і тоді оцінка буде точнішою, або розширювати довірчий інтервал. ◊

На практиці параметр θ невідомий і нам потрібно мати якийсь практичний метод підрахунку нижньої та верхньої меж $\theta_{\text{Н}}(s)$ та $\theta_{\text{В}}(s)$ так, щоб отримати необхідний довірчий рівень $1 - \alpha$. Зазвичай $1 - \alpha$ вибирають рівним 0,90; 0,95 або 0,99.

Для оцінок, що використовуються в теорії вибіркових обстежень, дуже важко вказати метод, що давав би точний довірчий рівень $1 - \alpha$. Тому в основному працюють з деякими наближеннями. Якщо $\widehat{\theta}$ – це оцінка параметра θ, то довірчий інтервал для θ з рівнем довіри, що приблизно дорівнює $1 - \alpha$, будують так:

$$\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathcal{D}}(\widehat{\theta})}, \tag{2.3}$$

де $z_{1-\alpha/2}$ – квантиль нормального розподілу рівня $1 - \alpha/2$.

Довірчий інтервал (2.3) буде містити невідоме значення параметра θ в $(1-\alpha) \times 100$ випадках із 100 при використанні вибіркового дизайну $p(s)$, якщо виконуються умови:

1) вибірковий розподіл оцінки $\widehat{\theta}$ приблизно нормальний із середнім θ та дисперсією $\mathcal{D}(\widehat{\theta})$;

2) існує строго конзистентна оцінка[2] $\widehat{\mathcal{D}}(\widehat{\theta})$ для дисперсії $\mathcal{D}(\widehat{\theta})$.

Тоді, при виконанні умови 1) $\frac{\widehat{\theta}-\theta}{\sqrt{\mathcal{D}(\widehat{\theta})}} \sim \mathcal{N}(0,1)$; умови 2) $\left(\frac{\mathcal{D}(\widehat{\theta})}{\widehat{\mathcal{D}}(\widehat{\theta})}\right)^{1/2} \approx 1$ з імовірністю приблизно рівною 1, коли n досить

[2]Поняття строгої конзистентності оцінки в теорії вибіркових обстежень дещо відрізняється від того, яким оперують у загальній теорії математичної статистики. Неможливість використання класичного поняття конзистентності пов'язана зі скінченним розміром генеральної сукупності. Для визначення цього поняття переходять до розгляду так званих «суперсукупностей» або «суперпопуляцій». Приклад застосування такого переходу див. в [16].

велике. Отже, в цьому випадку

$$\frac{\widehat{\theta}-\theta}{\sqrt{\widehat{\mathcal{D}}(\widehat{\theta})}}=\frac{\widehat{\theta}-\theta}{\sqrt{\mathcal{D}(\widehat{\theta})}}\left(\frac{\mathcal{D}(\widehat{\theta})}{\widehat{\mathcal{D}}(\widehat{\theta})}\right)^{1/2}\sim\mathcal{N}(0,1).$$

Зауваження 2.3. Якщо для побудови довірчого інтервалу використовувати значення $\mathcal{D}(\widehat{\theta})$, то довірчий рівень $1-\alpha$ буде досягатися точніше. Але, оскільки на практиці значення $\mathcal{D}(\widehat{\theta})$ невідоме, тому можливо лише користуватись оцінкою дисперсії $\widehat{\mathcal{D}}(\widehat{\theta})$.

Зауваження 2.4. Точність нормальної апроксимації в (2.3) залежить від набору значень характеристики y у генеральній сукупності. Якщо на діаграмі розсіювання ці значення розміщені асиметрично відносно середнього значення або містять нетипові одиниці, то для того, щоб отримати нормальну апроксимацію для довірчих інтервалів, потрібно мати досить великий розмір вибірки.

У деяких простих випадках можливо вказати, яким має бути середній розмір вибірки n, щоб отримати нормальну апроксимацію. Але в більшості випадків це зробити не так просто.

Є два способи перевірки того, чи досягається бажаний довірчий рівень $1-\alpha$ при побудові довірчого інтервалу за допомогою виразу (2.3) при заданому вибірковому плані $p(s)$, розмірі вибірки n_s та розподілі генеральної сукупності, – теоретичний та емпіричний.

Теоретичний полягає в перевірці того, чи є оцінка $\widehat{\theta}$ асимптотично нормальною при великих розмірах вибірки. У цьому напрямку отримано деякі результати [15]. Одним із важливих результатів є теорема Гаєка. У 1960 році цей чеський математик отримав асимптотичний розподіл для оцінки Горвіца–Томпсона при простому випадковому відборі без повернення. Дещо спрощено це твердження для оцінки сумарного виглядає так:

При деяких технічних умовах, та коли n, N, $N-n$ є «досить великими»

$$\frac{\widehat{t}_\pi-T}{\sqrt{N^2(1-\frac{n}{N})\frac{S^2}{n}}}\sim\mathcal{N}(0,1).$$

Але відповідь на питання, що більш за все цікавить практика: який розмір вибірки необхідний для нормальної апроксимації, такі результати дати не можуть.

Емпіричний полягає у використанні методу Монте-Карло. Серія з K вибірок (скажімо, $K=10\ 000$) заданого розміру отримується з сукупності, що є моделлю генеральної сукупності (нею може бути сукупність, інформація про яку була отримана з суцільного обстеження в попередні роки) відповідно до дизайну $p(s)$. Для кожної вибірки s підраховуються $\widehat{\theta}$, $\widehat{\mathcal{D}}(\widehat{\theta})$ та $DI(s)$, а потім перевіряється, чи належить θ отриманому довірчому інтервалу. Якщо R – це кількість інтервалів, що містять θ, то емпіричним довірчим рівнем буде величина $\frac{R}{K}\approx 1-\alpha$.

Подібні експерименти проводились для різних оцінок, вибіркових дизайнів, генеральних сукупностей та розмірів вибірок. Приклад такого експерименту наведено в [20, розділ 7.9.1].

2.5. Визначення розміру вибірки

Перед тим, як обчислити необхідний розмір вибірки, потрібно спочатку визначитись з такими параметрами: *допустима похибка* e та *рівень довіри* $1-\alpha$, що необхідні для кожного обстеження. Вони пов'язані співвідношенням

$$P(|\theta-\widehat{\theta}|\le e)=1-\alpha.$$

Ці дві величини є конкурентами. Чим більшою буде точність (значення похибки e – мале), тим меншим буде рівень довіри $1-\alpha$. Загальноприйнятими значеннями рівня довіри $1-\alpha$ є 0,90; 0,95; 0,99, для яких $\alpha=0,1;0,05;0,01$. Точність залежить від типу та мети обстеження. Якщо обстеження проводиться на регулярній основі з метою виявлення динаміки деякої характеристики, то точність повинна бути досить високою (похибка e мала). Якщо проводиться пробне обстеження, то висока точність не обов'язкова.

Розмір похибки e в основному залежить від дисперсії оцінки, яка у свою чергу буде зменшуватись зі збільшенням кількості елементів у вибірці. Але тут ще одним важливим фактором є

бюджет обстеження. Отримати велику вибірку, а отже і високу точність неможливо за умови обмеженого бюджету. Тому остаточно визначити розмір вибірки можна лише після врахування всіх цих факторів та знаходження деякого компромісу.

Визначення необхідного розміру вибірки продемонструємо на прикладі ПВВбП при оцінюванні середнього за допомогою π-оцінки $\widehat{\overline{y}}_\pi = \overline{y}$.

Приклад 2.3. Для простого випадкового відбору без повернення

$$e = z_{1-\alpha/2}\sqrt{\mathcal{D}(\overline{y})} = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}.$$

Звідси отримаємо

$$n = \frac{z^2_{1-\alpha/2}S^2}{e^2 + \frac{z^2_{1-\alpha/2}S^2}{N}}.$$

Оскільки абсолютне значення похибки залежить від одиниць розмірності, в яких вимірюється характеристика, тому частіше користуються безрозмірною величиною – відносною похибкою. Тоді необхідний розмір вибірки знаходять із рівності

$$P\left(\left|\frac{\theta-\widehat{\theta}}{\theta}\right| \le \widetilde{e}\right) = 1-\alpha.$$

У цьому випадку значення n при простому випадковому відборі без повернення буде залежати вже не від дисперсії в генеральній сукупності, а від коефіцієнта варіації $CV_y = \sqrt{S^2}/\overline{Y}$:

$$n = \frac{z^2_{1-\alpha/2}(CV_y)^2}{\widetilde{e}^2 + \frac{z^2_{1-\alpha/2}(CV_y)^2}{N}}.$$

◊

На практиці ні S^2, ні CV_y не відомі перед проведенням обстеження, тому ці значення зазвичай запозичують з подібних суцільних обстежень, або, якщо про генеральну сукупність немає жодної інформації, проводять пробне обстеження для визначення всіх необхідних для процесу планування параметрів.

2.6. Вправи та питання для самоконтролю

2.1. Що є π-оцінкою для середнього $\overline{Y}$ при ПВВбП?

2.2. Як буде змінюватись дисперсія оцінки Горвіца–Томпсона при збільшенні розміру вибірки n у випадку ПВВбП?

2.3. В яких випадках краще користуватись поняттям «відносна похибка», а в яких – просто «похибка»?

2.4. Метою обстеження є оцінка площі фермерських угідь у деякому районі Київської області. У цьому районі зареєстровано 2 015 фермерських господарств. Із них за допомогою простого випадкового відбору без повернення було відібрано 100 господарств та отримано інформацію про площу земель (y_k), що обробляється кожним із них:

$$\sum_{k\in s} y_k = 2\,856\text{ га} \quad \text{та} \quad \sum_{k\in s} y_k^2 = 156\,248\text{ га}^2.$$

1) Оцінити за допомогою оцінки Горвіца–Томпсона загальну площу землі, що обробляється фермерськими господарствами в цьому районі.

2) Побудувати 95-відсотковий довірчий інтервал для цієї величини.

2.5. Отримати вибірку розміру 4 з генеральної сукупності розміру 10 за допомогою схем відбору 1, 2, 3, що наведені в пункті 2.1. *Вказівка*: для генерування рівномірно розподілених на [0,1] чисел можна скористатись таблицею випадкових чисел із додатку 1.

2.6. Типове соціологічне опитування громадської думки робиться на основі вибірки, що складається приблизно з 1 000 осіб. Припустимо, що вибірка для такого обстеження отримана за допомогою простого випадкового відбору без повернення на основі списку всіх громадян України, включаючи Вас. За даними Держкомстату України населення країни станом на 2009 рік становило приблизно 46 млн осіб.

1) Яка ймовірність того, що Ви потрапите у вибірку тих, хто буде опитаний?

2) Припустимо, що таким чином утворено 2 000 незалежних вибірок. Яка ймовірність того, що Ви *не* потрапите в жодну з цих вибірок?

3) Скільки вибірок потрібно утворити для того, щоб з імовірністю 0,5 Ви потрапили хоча б в одну з них?

2.7. Проводиться обстеження працівників підприємства з метою визначення пропорції тих, хто хворіє на професійну хворобу. На підприємстві працює 2 000 працівників. Із зовнішніх джерел відомо, що зазвичай на подібних підприємствах професійну хворобу мають 3 з 10 працівників. При обстеженні планується використання дизайну ПВВбП з допустимою похибкою 0,01 та рівнем довіри 95 %.

1) Яким має бути необхідний розмір вибірки n?

2) Якщо немає ніякої інформації про подібні захворювання, яким повинен бути розмір вибірки n?

2.8. Потрібно оцінити кількість людей серед населення України, що хворіють на туберкульоз. При обстеженні планується використання ПВВбП. Якщо реальна частка хворих на туберкульоз складає 0,01 %, скільки осіб потрібно обстежити, щоб коефіцієнт варіації оцінки $CV(\widehat{P}) = \frac{\sqrt{\mathcal{D}(\widehat{P})}}{E\widehat{P}}$ не перевищував 5 %? Проаналізувати отриманий результат.

2.9. При обстеженні жителів деякого великого міста будуть досліджуватись дві величини:

P_1 – частка тих, хто має пральну машину;

P_2 – частка тих, хто має доступ до мережі інтернет.

Із достовірних джерел відомо, що $40\ \% \le P_1 \le 60\ \%$ та $5\ \% \le P_2 \le 15\ \%$. Яким має бути розмір вибірки при застосуванні ПВВбП, якщо ми хочемо оцінити *одночасно* параметр P_1 з точністю ± 2 % та параметр P_2 з точністю ± 1 % при рівні довіри 95 %?

2.10. На виборах в останньому турі потрібно вибрати одного з двох кандидатів (третього варіанту немає). Напередодні дня виборів проведено опитування громадської думки з метою визначення

переможця. Нехай n – кількість тих, хто був опитаний при застосуванні ПВВбП (припускається, що $n > 100$ та розмір генеральної сукупності великий порівняно з розміром вибірки).

Якою повинна бути різниця між відсотками голосів, набраних кандидатами, для того, щоб у результаті опитування можна було з імовірністю 0,95 визначити переможця виборів? Обчислити ці значення для різних n.

Вказівка: якщо припустити, що кандидат A програє вибори та P_A – це відсоток голосів, які він отримає в день виборів, то нехай $\widehat{P}_A$ – це відсоток голосів, що отримав кандидат A напередодні під час опитування. Тоді задача буде зводитись до знаходження критичної області для $\widehat{P}_A$, для якої ймовірність визнання кандидата A переможцем напередодні виборів буде складати менше 0,05. Тобто, потрібно знайти таке значення c, щоб

$$P\{\widehat{P}_A > c | P_A < 50\ \%\} \le 0,05.$$

Розділ 3

Відбір Бернуллі

3.1. Основні властивості

Відбір Бернуллі є надзвичайно простим видом випадкового відбору. Ми його використовували для ілюстрації деяких теоретичних результатів (див. приклади 1.2, 1.4, 1.9, 1.10).

Отже, при відборі Бернуллі індикатори включення $I_1, I_2, ..., I_N$ є незалежними і однаково розподіленими випадковими величинами. Для всіх елементів імовірність їх включення однакова, тобто $\pi_k = \pi$ для всіх k та

$$P(I_k = 1) = \pi, \quad P(I_k = 0) = 1 - \pi.$$

Вибірковий дизайн

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s},$$

де n_s – розмір вибірки, що є біноміально розподіленою випадковою величиною з

$$E_{\text{ВБ}} n_s = N\pi, \quad \mathcal{D}_{\text{ВБ}} n_s = N\pi(1-\pi).$$

Розкид можливих значень випадкової величини n_s можна оцінити, якщо наблизити біноміальний розподіл нормальним

$$DI(n_s) = N\pi \pm z_{1-\alpha/2}\sqrt{N\pi(1-\pi)}$$

з рівнем довіри приблизно рівним $1-\alpha$.

Визначимо ймовірності включення

$$\begin{aligned} \pi_k &= \pi \quad \forall k; \\ \pi_{kl} &= \pi^2 \quad \forall k \neq l; \\ \pi_{kk} &= \pi \quad \forall k \end{aligned}$$

та

$$\begin{aligned} \Delta_{kl} &= \pi_{kl} - \pi_k\pi_l = \pi^2 - \pi^2 = 0; \\ \Delta_{kk} &= \pi - \pi^2 = \pi(1-\pi). \end{aligned}$$

3.2. Оцінка Горвіца–Томпсона при відборі Бернуллі

***Твердження* 3.1.** *При вибірковому дизайні Бернуллі оцінки Горвіца–Томпсона для сумарного T та середнього $\overline{Y}$ набувають вигляду*

$$\widehat{t}_\pi = \frac{1}{\pi}\sum_{k\in s} y_k, \quad \widehat{\overline{y}}_\pi = \frac{1}{N\pi}\sum_{k\in s} y_k$$

з дисперсіями

$$\begin{aligned} \mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi) &= \left(\frac{1}{\pi} - 1\right)\sum_{k\in U} y_k^2 = \frac{1-\pi}{\pi}\sum_{k\in U} y_k^2, \\ \mathcal{D}_{\text{ВБ}}(\widehat{\overline{y}}_\pi) &= \frac{1-\pi}{N^2\pi}\sum_{k\in U} y_k^2 \end{aligned}$$

та оцінками для дисперсій

$$\begin{aligned} \widehat{\mathcal{D}}_{\text{ВБ}}(\widehat{t}_\pi) &= \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_{k\in s} y_k^2 = \frac{1-\pi}{\pi^2}\sum_{k\in s} y_k^2, \\ \widehat{\mathcal{D}}_{\text{ВБ}}(\widehat{\overline{y}}_\pi) &= \frac{1-\pi}{N^2\pi^2}\sum_{k\in s} y_k^2. \end{aligned}$$

Доведення. Справді,

$$\widehat{t}_\pi = \sum_{k\in s}\frac{y_k}{\pi_k} = \frac{1}{\pi}\sum_{k\in s} y_k,$$

та

$$\begin{aligned} \mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi) &= \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_\pi}{\pi_\pi}\frac{y_l}{\pi_l} = \sum_{k\in U}\Delta_{kk}\left(\frac{y_k}{\pi_k}\right)^2 = \\ &= \pi(1-\pi)\frac{1}{\pi^2}\sum_{k\in U} y_k^2 = \frac{1-\pi}{\pi}\sum_{k\in U} y_k^2 \end{aligned}$$

з незміщеною оцінкою

$$\widehat{\mathcal{D}}_{\text{ВБ}}(\widehat{t}_\pi) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\cdot\frac{y_\pi}{\pi_\pi}\cdot\frac{y_l}{\pi_l} = \sum_{k\in s}\frac{\Delta_{kk}}{\pi_{kk}}\left(\frac{y_k}{\pi_k}\right)^2 =$$
$$= \frac{\pi(1-\pi)}{\pi}\frac{1}{\pi^2}\sum_{k\in s}y_k^2 = \frac{1-\pi}{\pi^2}\sum_{k\in s}y_k^2.$$

Відповідні вирази для середнього отримуємо, враховуючи, що $\widehat{\overline{y}}_\pi = \frac{1}{N}\widehat{t}_\pi$ та $\mathcal{D}(\widehat{\overline{y}}_\pi) = \frac{1}{N^2}\mathcal{D}(\widehat{t}_\pi)$. □

***Твердження* 3.2.** *При застосуванні відбору Бернуллі оцінки для кількості N_d елементів з деякої підсукупності U_d та частки цих елементів у сукупності P_d набувають вигляду*

$$\widehat{N}_d = \frac{n_{s_d}}{\pi};\quad \widehat{P}_d = \frac{n_{s_d}}{N\pi},$$

де $s_d = s\cap U_d$, n_{s_d} – кількість елементів у вибірці, що належать підсукупності U_d. Запишемо дисперсії цих оцінок

$$\mathcal{D}_{\text{ВБ}}(\widehat{N}_d) = \frac{1-\pi}{\pi}N_d,\quad \mathcal{D}_{\text{ВБ}}(\widehat{P}_d) = \frac{1-\pi}{N^2\pi}N_d$$

та оцінки для дисперсій

$$\widehat{\mathcal{D}}_{\text{ВБ}}(\widehat{N}_d) = \frac{1-\pi}{\pi^2}n_{s_d},\quad \widehat{\mathcal{D}}_{\text{ВБ}}(\widehat{P}_d) = \frac{1-\pi}{N^2\pi^2}n_{s_d}.$$

Доведення. Дане твердження отримується з попереднього, враховуючи, що $N_d = \sum_{k\in U} z_{dk}$, де $z_{dk} = 1$, якщо k-й елемент належить підсукупності U_d, та 0 – в іншому випадку, а $P_d = \frac{N_d}{N}$. □

Приклад 3.1. Застосуємо отримані результати для дослідження генеральної сукупності, що складається з 4 елементів. Припустимо, що значення характеристики y нам відомі для кожного елемента (табл. 4).

Таблиця 4. **Елементи генеральної сукупності з відповідними значеннями характеристики y**

Елемент	1	2	3	4
y_k	2	6	8	10

Застосовуючи відбір Бернуллі, можна отримати 2^4 різних вибірок, включаючи порожню множину та всю генеральну сукупність (табл. 5).

Таблиця 5. **16 можливих вибірок при відборі Бернуллі та значення $\widehat{\overline{y}}_\pi$ при $\pi = 1/2$**

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Елементи	∅	1	2	3	4	1 2	1 3	1 4	2 3	2 4	3 4	1 2 3	1 2 4	1 3 4	2 3 4	1 2 3 4
Значення	-	2	6	8	10	2 6	2 8	2 10	6 8	6 10	8 10	2 6 8	2 6 10	2 8 10	6 8 10	2 6 8 10
$\widehat{\overline{y}}_\pi$	0	1	3	4	5	4	5	6	7	8	9	8	9	10	12	13

Запишемо вибірковий розподіл оцінки (табл. 6).

Таблиця 6. **Вибірковий розподіл $\widehat{\overline{y}}_\pi$**

$\widehat{\overline{y}}_\pi$	0	1	3	4	5	6	7	8	9	10	12	13
P	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Застосування оцінки Горвіца–Томпсона при відборі Бернуллі не є ефективним. Як ми бачимо з вибіркового розподілу, можливі значення оцінки $\widehat{\overline{y}}_\pi$ досить сильно розкидані відносно реального значення середнього $\overline{Y}$ =6,5, а значення 6 та 7, що є найближчими до 6,5, можуть бути отримані з такою самою ймовірністю, як і значення 0 або 13. Більш того, при обстеженні всієї генеральної сукупності (вибірка 16) отримуються значення, що дуже далеке від $\overline{Y}$, хоча в цьому випадку ми вже можемо обчислити реальне значення.

Значення дисперсії оцінки $\mathcal{D}_{\text{ВБ}}(\widehat{\overline{y}}_\pi) = 25,5$ та коефіцієнта варіації $CV(\widehat{\overline{y}}_\pi) = \frac{\sqrt{\mathcal{D}_{\text{ВБ}}(\widehat{\overline{y}}_\pi)}}{\overline{Y}} = 0,77$ (77 %) є також досить великими, основною причиною чого є змінний розмір вибірок при відборі Бернуллі. ◊

3.3. Недоліки відбору Бернуллі

Основними недоліками відбору Бернуллі є:

- більша дисперсія оцінки Горвіца–Томпсона, ніж при ПВВбП;
- змінний розмір вибірки, що викликає труднощі при плануванні.

З першим недоліком можна боротися за допомогою використання іншої оцінки. Якщо ми зафіксуємо середній очікуваний розмір вибірки $En_s = N\pi = n$, то $\pi = \frac{n}{N}$ і

$$\widehat{t}_\pi = \frac{N}{n}\sum_{k\in s} y_k = N\left(\frac{1}{n}\sum_{k\in s} y_k\right).$$

Величина в дужках – є вибірковим середнім, значення якого поширюється на всі N елементів генеральної сукупності. Але підсумування проводиться по всіх елементах, що потрапили у вибірку, в той час як усереднення – по середньому очікуваному розміру вибірки n. Логічніше було б усереднювати по тій кількості елементів, що реально потрапляють у вибірку. У цьому випадку отримаємо альтернативну оцінку

$$\widehat{t}_{\text{альт}} = \frac{N}{n_s}\sum_{k\in s} y_k = \frac{n_s}{n}\widehat{t}_\pi. \tag{3.1}$$

Випадковий розмір n_s в (3.1) зменшує варіацію, головною причиною якої був змінний розмір вибірки. Деякими спеціальними методами можна отримати наближений вираз для дисперсії (див. підрозділ 9.3.):

$$\mathcal{D}_{\text{ВБ}}(\widehat{t}_{\text{альт}}) \approx N\left(\frac{1}{\pi} - 1\right)S^2 = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2 = N^2\frac{1-f}{n}S^2.$$

Тоді

$$\frac{\mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi)}{\mathcal{D}_{\text{ВБ}}(\widehat{t}_{\text{альт}})} \approx \left(1 + \frac{1}{(CV_U)^2}\right).$$

Ефективність використання оцінки $\widehat{t}_{\text{альт}}$ особливо відчутна, коли коефіцієнт варіації генеральної сукупності CV_U малий, тобто коли генеральна сукупність досить однорідна. Таким чином, використання іншої (альтернативної) оцінки усуває перший недолік відбору Бернуллі – велику дисперсію:

$$\mathcal{D}_{\text{ВБ}}(\widehat{t}_{\text{альт}}) \approx \mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi).$$

Але інший недолік залишається. А при обмеженому бюджеті це може бути дуже суттєво.

І хоч на практиці всі намагаються уникати вибіркових дизайнів зі змінним розміром вибірки, за присутності невідповідей такі вибіркові дизайни, як ВБ, можуть служити моделлю отримання відповідей.

Крім того, у випадках, коли потрібно оцінити не лише загальний параметр для деякої сукупності, але й знайти відповідні оцінки для підсукупностей, контролювати розміри вибірок у кожній підсукупності дуже важко. І тоді такого типу відбори стають у нагоді.

3.4. Дизайн-ефект

У теорії вибіркових обстежень простий випадковий відбір без повернення відіграє дуже важливу роль. Хоч у чистому вигляді цей відбір рідко використовується на практиці, але він є найпростішим і служить основою для більш складних відборів. Тому йому надали роль еталона, з яким порівнюють усі інші відбори.

При простому випадковому відборі без повернення оцінка Горвіца–Томпсона $\widehat{t}_\pi = N\overline{y}$ має дисперсію

$$\mathcal{D}_{\text{ПВВбП}}\left(\widehat{t}_\pi\right) = N^2\frac{1-f}{n}S^2 = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2, \quad f = \frac{n}{N}.$$

Нехай $p(s)$ – деякий відмінний від ПВВбП вибірковий дизайн з таким самим середнім розміром вибірки $En_s = \sum\limits_{k\in U}\pi_k = n$ (це

потрібно, щоб порівняння було справедливим). Припустимо, що π-оцінку $\widehat{t}_\pi$ можна використати і для цього вибіркового дизайну. Тоді *дизайн-ефектом* називається величина

$$deff\left(p(s),\widehat{t}_\pi\right) := \frac{\mathcal{D}_p(\widehat{t}_\pi)}{\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)} = \frac{\sum\sum_{k,l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}}{N^2(\frac{1}{n}-\frac{1}{N})S^2}.$$

Величина $deff$ визначає ефективність стратегії, що складається з вибіркового плану $p(s)$ та оцінки $\widehat{t}_\pi$, порівняно з простим випадковим відбором та оцінкою $N\overline{y}$.

Якщо $deff\left(p,\widehat{t}_\pi\right) > 1$, то точність втрачається при застосуванні дизайну $p(s)$ для оцінки $\widehat{t}_\pi$. Якщо $deff\left(p,\widehat{t}_\pi\right) < 1$, то отримуємо точніші результати порівняно з ПВВбП.

Оскільки в обох випадках використовується оцінка Горвіца–Томпсона $\widehat{t}_\pi$, то причиною відмінностей є вибірковий дизайн $p(s)$, що впливає на розподіл $\widehat{t}_\pi$.

Приклад 3.2. Дослідимо, який дизайн-ефект має відбір Бернуллі. Нагадаємо, що в цьому випадку

$$\widehat{t}_\pi = \sum_{k\in s}\frac{y_k}{\pi_k} = \frac{1}{\pi}\sum_{k\in s} y_k$$

та

$$\begin{aligned}\mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi) &= \frac{1-\pi}{\pi}\sum_{k\in U}y_k^2 = \frac{1-\pi}{\pi}\left[(N-1)S^2+N(\overline{Y})^2\right] = \\ &= \frac{1-\pi}{\pi}NS^2\left[1-\frac{1}{N}+\left(\frac{\overline{Y}}{S}\right)^2\right] = \\ &= \frac{1-\pi}{\pi}\left(1-\frac{1}{N}+\frac{1}{(CV_U)^2}\right)NS^2,\end{aligned}$$

де $CV_U = \frac{S}{\overline{Y}}$ – коефіцієнт варіації генеральної сукупності.

Для того, щоб порівняння було справедливим, зафіксуємо середній розмір вибірки $En_s = N\pi = n$, тобто $\pi = \frac{n}{N}$. Тоді

$$\begin{aligned}deff(\text{ВБ},\widehat{t}_\pi) &= \frac{\mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi)}{\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)} = \\ &= \frac{\left(1-\frac{n}{N}\right)\frac{N}{n}NS^2\left(1-\frac{1}{N}+\frac{1}{(CV_U)^2}\right)}{N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}} = \\ &= \left(1-\frac{1}{N}+\frac{1}{(CV_U)^2}\right)\approx 1+\frac{1}{(CV_U)^2}.\end{aligned}$$

Для більшості генеральних сукупностей коефіцієнт варіації лежить у межах: $0,5 \le CV_U \le 1$. Якщо $CV_U = 1$, то $deff(\text{ВБ},\widehat{t}_\pi) = 2$, якщо $CV_U = 0,5$, то $deff(\text{ВБ},\widehat{t}_\pi) = 5$. Отже, в більшості випадків відбір Бернуллі є менш ефективним (точним), ніж простий випадковий відбір. Це пояснюється тим, що вибірковий дизайн ВБ має змінний розмір вибірок. ◊

Приклад 3.3. Для генеральної сукупності з прикладу 2.1 коефіцієнт варіації

$$CV_U = \frac{S}{\overline{Y}} = \frac{\sqrt{64/5}}{8} = \frac{1}{\sqrt{5}} \approx 0{,}45, \text{ тобто } deff(\text{ВБ},\widehat{t}_\pi) = 6.$$

Отже, можна зробити висновок, що використання відбору Бернуллі буде в 6 разів менш ефективним. Це означає, що при однаковому середньому розмірі вибірки оцінки при відборі Бернуллі будуть мати дисперсію в 6 разів більшу ніж при ПВВбП. Або, що еквівалентно, для отримання однакової точності при відборі Бернуллі та простому випадковому відборі потрібно в першому випадку обстежити в середньому в 6 разів більше елементів. ◊

3.5. Вправи та питання для самоконтролю

3.1. Чи можна визначити відбір Бернуллі, однозначно зафіксувавши середній розмір вибірки?

3.2. Які основні недоліки відбору Бернуллі? Чи можна їх усунути?

3.3. Які властивості повинна мати генеральна сукупність для того, щоб дизайн-ефект відбору Бернуллі приблизно дорівнював 1?

3.4. Нехай генеральна сукупність U поділена на три неперетинні підсукупності U_1, U_2 та U_3, що мають розміри $N_1 = 300$, $N_2 = 200$ та $N_3 = 100$. Для кожного елемента k включення або невключення у вибірку s визначається в результаті випробувань Бернуллі, при яких елемент k потрапляє у вибірку з імовірністю π_k. Усі випробування незалежні.

1) Якщо π_k =0,1 для $k \in U_1$, π_k =0,3 для $k \in U_2$ та π_k =0,7 для $k \in U_3$, яким буде математичне сподівання та дисперсія величини n_s – розміру вибірки, при такому вибірковому дизайні?

2) Якщо припустити, що всі π_k однакові для всієї генеральної сукупності, яким має бути це значення для того, щоб отримати таке саме математичне сподівання величини n_s, як і в попередньому пункті? Підрахувати значення дисперсії розміру вибірки n_s в цьому випадку.

3.5. Нехай s – це вибірка, що була отримана в результаті відбору Бернуллі з генеральної сукупності U розміру N якщо для всіх $k \in U$ $\pi_k = \pi$. Нехай n_s – це випадковий розмір вибірки s. Довести, що умовна ймовірність отримання вибірки s за умови фіксованого розміру n_s співпадає з імовірністю отримання цієї вибірки при простому випадковому відборі без повернення, коли $n = n_s$.

3.6. Для виявлення сумарної кількості дітей віком до 2-х років, яким не була вчасно зроблена вакцинація, проводиться вибіркове обстеження в деякій області України. Відомо, що в цій області є 120 медичних пунктів, де проводиться вакцинація. Характеристикою y_k, що вивчається, є кількість дітей, яким не були вчасно зроблені щеплення в медпункті k. Для обстеження планується використання відбору Бернуллі. Відомо, що $\sum_{k\in U} y_k \approx 2\,400$ та $\sum_{k\in U} y_k^2 \approx 49\,000$. Якою має бути середня кількість медпунктів, що підлягають обстеженню, при такому відборі та застосуванні

1) оцінки Горвіца–Томпсона $\widehat{t}_\pi$;

2) альтернативної оцінки $\widehat{t}_{\text{альт}}$, що визначена формулою (3.1),

якщо коефіцієнт варіації кожної з цих оцінок не повинен перевищувати 10 %?

3.7. До деякого населеного пункту України належать 2 500 домогосподарств. Вибірка розміру $n_s = 900$ була отримана в результаті застосування відбору Бернуллі з $\pi = 0,5$ з метою визначення частки тих домогосподарств, що перебувають за межею бідності. У результаті обстеження було виявлено, що таких домогосподарств у вибірці 350. Потрібно оцінити відсоток тих домогосподарств, що перебувають за межею бідності в цьому населеному пункті, та побудувати 95-відсотковий довірчий інтервал для цієї оцінки.

3.8. Для генеральної сукупності U розміру $N = 10\,000$ коефіцієнт варіації $CV_U = 0,534$. Визначити необхідний (середній) розмір вибірки для оцінки середнього деякої характеристики y так, щоб коефіцієнт варіації оцінки Горвіца–Томпсона $CV(\widehat{\overline{y}}_\pi) = \sqrt{\mathcal{D}(\widehat{\overline{y}}_\pi)}/\overline{Y}$ не перевищував 1 % при застосуванні:

1) простого випадкового відбору без повернення;

2) відбору Бернуллі.

Яким повинен бути середній розмір вибірки при використанні альтернативної оцінки (3.1) для відбору Бернуллі?

Розділ 4

Систематичний відбір

4.1. Основні поняття та результати

Головною перевагою систематичного відбору є простота його використання на практиці. Нехай N – розмір генеральної сукупності U, $a \in \mathbb{N}$ – деяке фіксоване число. Перший елемент вибірки вибирається випадковим чином серед перших a елементів сукупності. Обране таким чином число r, $1 \leq r \leq a$, називається *випадковим стартом (випадковим початком)*, а число a – *вибірковим інтервалом*. Кожен елемент $1, 2, \ldots, a$ має однакову ймовірність бути обраним: $\frac{1}{a}$. Далі у вибірку потрапляють елементи з кроком a, тобто r, $r+a$, $r+2a$, $\ldots$

Таким чином можна отримати лише a різних вибірок, кожна з яких має однакову ймовірність бути обраною: $\frac{1}{a}$.

Нехай $n = \left[\frac{N}{a}\right]$, тоді $N = an + c$, де $0 \leq c < a$.

Якщо $c = 0$, то розмір усіх систематичних вибірок дорівнюватиме n. Якщо $c \neq 0$, то систематичні вибірки можуть мати різні розміри: n чи $n+1$.

Позначимо $\mathfrak{F}_{\text{СВ}} = \{s_1, \ldots, s_a\}$ множину всіх можливих вибірок при систематичному відборі. Тоді

$$p(s) = \begin{cases} \frac{1}{a}, & \text{якщо } s \in \mathfrak{F}_{\text{СВ}}; \\ 0, & \text{в іншому випадку.} \end{cases}$$

Для того, щоб визначити, якого вигляду набуває оцінка Горвіца–Томпсона та її дисперсія при систематичному відборі, потрібно знайти ймовірності включення першого та другого порядків:

$$\pi_k = \frac{1}{a}, \quad \forall k;$$

$$\pi_{kl} = \begin{cases} \frac{1}{a}, & \text{якщо } k \text{ та } l \text{ належать одній вибірці } s; \\ 0, & \text{якщо ні,} \end{cases} \quad \forall k \neq l.$$

Таким чином, дизайн при систематичному відборі *не є вимірним* (тобто, не для всіх $k, l \in U$ $\pi_{kl} > 0$).

Оскільки вибірки $s_1, \ldots, s_a$ не перетинаються та $\bigcup_{r=1}^{a} s_r = U$, то сумарне T ми можемо записати як

$$T = \sum_{r=1}^{a} t_{s_r}, \quad \text{де} \quad t_{s_r} = \sum_{k \in s_r} y_k.$$

Систематичний відбір можна розглядати як відбір однієї з a підсукупностей генеральної сукупності з рівними ймовірностями $\frac{1}{a}$ з подальшим обстеженням всіх елементів вибраної підсукупності.

***Твердження* 4.1.** *При систематичному відборі з вибірковим інтервалом a, π-оцінка для сумарного T має вигляд*

$$\widehat{t}_\pi = a \cdot t_s, \quad \textit{де} \quad t_s = \sum_{k \in s} y_k, \quad s \in \mathfrak{F}_{\text{СВ}},$$

з дисперсією

$$\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right) = a \cdot (a-1)\, S_t^2, \tag{4.1}$$

де

$$S_t^2 = \frac{1}{a-1} \sum_{r=1}^{a} \left(t_{s_r} - \bar{t}\right)^2, \quad \bar{t} = \frac{1}{a} \cdot T = \frac{1}{a} \sum_{r=1}^{a} t_{s_r}.$$

При цьому дисперсія $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$ буде малою, якщо сумарні t_r по вибірках s_r будуть приблизно рівними.

Доведення. Оскільки для всіх $k = \overline{1, N}$ $\pi_k = \frac{1}{a}$, то

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = a \sum_{k \in s} y_k = a t_s, \quad s \in \mathfrak{F}_{\text{СВ}}.$$

Тоді

$$\begin{aligned} \mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right) &= \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \\ &= \sum_{k \in U} \sum_{l \in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1\right) y_k y_l = \end{aligned}$$

$$= \sum_{k\in U}\sum_{l\in U}\frac{\pi_{kl}}{\pi_k\pi_l}y_k y_l - \left(\sum_{k\in U} y_k\right)^2 =$$

$$= a\sum_{r=1}^{a}\left(\sum_{k\in s_r}\sum_{l\in s_r} y_k y_l\right) - T^2 = a\sum_{r=1}^{a} t_{s_r}^2 - T^2 =$$

$$= a\left(\sum_{r=1}^{a} t_{s_r}^2 - a\bar{t}^2\right) = a\sum_{r=1}^{a}\left(t_{s_r} - \bar{t}\right)^2.$$

□

Зауваження 4.1. Оскільки вибірковий дизайн СВ *не є вимірним* (не для всіх $k, l \in U$ $\pi_{kl} > 0$), то використовувати вираз для оцінки дисперсії $\widehat{\mathcal{D}}\left(\widehat{t}_\pi\right)$ (формула (1.6)) *не можна*!!!

4.2. Розмір вибірки при систематичному відборі

При застосуванні систематичного відбору дуже важливим є контроль розміру n можливих вибірок. Нехай

$$N = an + c, \quad 0 \le c < a.$$

Якщо $c = 0$, то всі a можливих вибірок мають однаковий розмір n; якщо $c > 0$, то розмір вибірки дорівнює $n + 1$ при $r \le c$ та n при $r > c$.

Вимога, щоб a було цілим числом приводить у деяких ситуаціях до розміру вибірки, що є далеким від бажаного.

Приклад 4.1. Нехай $N = 149$ і бажаний розмір вибірки $n = 60$. Тоді, якщо вибірковий інтервал $a = 2$, то ми отримаємо розмір вибірки 74 або 75; при $a = 3$ – або 49, або 50. Отримати вибірку розміру 60 неможливо. ◊

Зауважимо, що коли n мале порівняно з N, то проблема не є настільки серйозною. Наведемо два способи її вирішення.

Метод дробового вибіркового інтервалу. Цей метод допускає, щоб a було нецілим числом. Нехай $a = \frac{N}{n}$, де n – бажаний

розмір вибірки. Згенеруємо випадкове число ξ, що має рівномірний розподіл на $[0, a]$. Вибірка в цьому випадку буде складатись з елементів $k \in U$, для яких

$$k - 1 < \xi + (j-1)\,a \le k, \quad j = 1, \ldots, n.$$

Або, це еквівалентно вибору з імовірністю $\frac{1}{N}$ випадкового цілого числа r, $1 \le r \le N$, та відбору у вибірку елементів k, для яких

$$(k-1)\,n < r + (j-1)\,N \le kn, \quad j = 1, \ldots, n.$$

Наприклад, елемент $k = 1$ буде відібраний якщо $0 < r \le n$. Ця подія наступає з імовірністю $\frac{n}{N}$. При цьому всі можливі вибірки мають розмір n.

Циклічний метод. У цьому випадку вибіркову структуру роблять циклічною, тобто після N-го елемента слідує 1-й і т. д. Вибирається випадкове число r $(1 \le r \le N)$. Нехай a – це найближче ціле число біля $\frac{N}{n}$. Тоді у вибірку потрапляють ті елементи k, для яких

$$\begin{array}{ll} k = r + (j-1)\,a, & \text{якщо } r + (j-1)\,a \le N, \\ k = r + (j-1)\,a - N, & \text{якщо } r + (j-1)\,a > N, \end{array} \quad j = 1, \ldots, n.$$

У цьому випадку розмір n буде однаковим для всіх вибірок.

Щоб підрахувати $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$ для обох цих методів, потрібно знайти ймовірності включення другого порядку π_{kl}, $k \ne l$. Але в цих випадках вибірки s_r не обов'язково неперетинні, тому при підрахунку ймовірностей включення потрібно цей факт враховувати.

У разі, коли $c = 0$, всі методи систематичного відбору еквівалентні. Якщо N досить велике відносно n, то різниця між цими методами буде незначною.

4.3. Ефективність систематичного відбору

Значення $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$, а отже, і ефективність систематичного відбору залежить від того, яким чином елементи генеральної сукупності впорядковані. Дисперсія буде малою якщо сумарні t_{s_r}

по кожній можливій вибірці будуть приблизно рівними. В якому випадку це може бути?

Для простоти розглянемо випадок, коли $N = an$, де a – ціле число. Тоді $a = \frac{N}{n}$ і

$$\begin{aligned} \widehat{t}_\pi &= \frac{N}{n}\sum_{k\in s_r} y_k = N\overline{y}_{s_r}, \\ \mathcal{D}_{\mathrm{CB}}\left(\widehat{t}_\pi\right) &= N^2\frac{1}{a}\sum_{r=1}^{a}\left(\overline{y}_{s_r}-\overline{Y}\right)^2 = Nn\sum_{r=1}^{a}\left(\overline{y}_{s_r}-\overline{Y}\right)^2. \end{aligned}$$

Розглянемо варіацію елементів генеральної сукупності, тобто

$$\begin{aligned} \sum_{k\in U}\left(y_k-\overline{Y}\right)^2 &= \sum_{r=1}^{a}\sum_{k\in s_r}\left(y_k-\overline{Y}\right)^2 = \\ &= \sum_{r=1}^{a}\sum_{k\in s_r}\left(y_k\pm\overline{y}_{s_r}-\overline{Y}\right)^2 = \\ &= \sum_{r=1}^{a}\sum_{k\in s_r}\left(y_k-\overline{y}_{s_r}\right)^2+\sum_{r=1}^{a}n\left(\overline{y}_{s_r}-\overline{Y}\right)^2. \end{aligned}$$

Отримуємо основну тотожність дисперсійного аналізу

$$SST = SSW + SSB,$$

де

$$\begin{aligned} SST &= \sum_{k\in U}\left(y_k-\overline{Y}\right)^2 \text{ – загальна варіація,} \\ SSW &= \sum_{r=1}^{a}\sum_{k\in s_r}\left(y_k-\overline{y}_{s_r}\right)^2 \text{ – внутрішньогрупова варіація,} \\ SSB &= n\sum_{r=1}^{a}\left(\overline{y}_{s_r}-\overline{Y}\right)^2 \text{ – міжгрупова варіація.} \end{aligned}$$

У цьому записі:

- SS від англ. `sum of squares`, що означає – *сума квадратів*;

- T від англ. `total` – означає *загальний*;
- W від англ. `within` – означає *всередині*;
- B від англ. `between` – означає *між*.

Зокрема, $S^2 = \frac{1}{N-1}SST$. У цьому випадку основний вклад у дисперсію оцінки робить міжгрупова дисперсія генеральної сукупності:

$$\mathcal{D}_{\mathrm{CB}}\left(\widehat{t}_\pi\right) = N\cdot SSB.$$

Тому, щоб дисперсія оцінки $\mathcal{D}_{\mathrm{CB}}\left(\widehat{t}_\pi\right)$ була меншою, бажано, щоб SSB була меншою, що означає, що SSW має бути по можливості більшою. Іншими словами, елементи всередині кожної групи повинні бути по можливості неоднорідними.

4.4. Міри однорідності

Однією з поширених мір однорідності в групі є *внутрішньогруповий коефіцієнт кореляції*:

$$\rho = 1-\frac{n}{n-1}\frac{SSW}{SST} = \frac{2\sum_{r=1}^{a}\left[\sum_{k<l}\sum_{s_r}\left(y_k-\overline{Y}\right)\left(y_l-\overline{Y}\right)\right]}{(n-1)(N-1)S^2}.$$

По суті, ρ – це міра кореляції між парами елементів всередині групи. Дослідимо, в яких випадках цей коефіцієнт набуває свого максимального та мінімального значень:

$$\rho_{\max} = 1 \Rightarrow SSW = 0,\ SST = SSB,$$

тобто, всі елементи всередині кожної групи однакові;

$$\rho_{\min} = 1-\frac{n}{n-1} = -\frac{1}{n-1} \Rightarrow SSB = 0,\ SSW = SST,$$

тобто, середні значення в кожній групі однакові.

Проілюструємо це на прикладі.

Приклад 4.2. Нехай генеральна сукупність $U = \{1, 1, 2, 2\}$. Розіб'ємо цю сукупність на групи (вибірки): $s_1 = \{1, 2\}$, $s_2 = \{1, 2\}$. Це означає, що ми застосували систематичний відбір з вибірковим інтервалом $a = 2$. Тоді з $N = 4$ елементів ми утворили вибірки розміру $n = 2$. Вибірка s_1 утворюється, коли випадковим стартом є $r = 1$, а s_2 – коли випадковим стартом є $r = 2$. Тоді

$$\bar{y}_1 = 1,5; \quad \bar{y}_2 = 1,5; \quad \overline{Y} = 1,5,$$

отже,

$$\begin{aligned} SSW &= 2 \cdot (1-1,5)^2 + 2 \cdot (2-1,5)^2 = 4 \cdot 0,25 = 1; \\ SSB &= 0; \\ SST &= 2 \cdot (1-1,5)^2 + 2 \cdot (2-1,5)^2 = 4 \cdot 0,25 = 1, \end{aligned}$$

і виконується основна тотожність дисперсійного аналізу $SST = SSW + SSB$, $1 = 1 + 0$. У цьому випадку $\rho = -1$, тобто елементи всередині групи неоднорідні, але середні по групах однорідні.

Якщо ж генеральна сукупність упорядкована іншим чином $U = \{1, 2, 1, 2\}$, то в результаті застосування того самого систематичного відбору отримаємо інше розбиття на групи (вибірки): $s_1 = \{1, 1\}$, $s_2 = \{2, 2\}$. У цьому випадку

$$\bar{y}_1 = 1, \quad \bar{y}_2 = 2, \quad \overline{Y} = 1,5,$$

$$\begin{aligned} SSW &= 0; \\ SSB &= 2 \cdot \left[(1-1,5)^2 + (2-1,5)^2\right] = 1; \\ SST &= 1. \end{aligned}$$

$\rho = 1$, а це означає, що елементи всередині групи однорідні, варіація елементів виникає за рахунок міжгрупової варіації SSB (варіації середніх по групах). ◊

Основний недолік коефіцієнта ρ полягає в тому, що його можна використовувати лише для груп однакового розміру. Тому

частіше віддають перевагу іншому коефіцієнту:

$$\delta = 1 - \frac{N-1}{N-a}\frac{SSW}{SST}. \tag{4.2}$$

Позначимо S_W^2 внутрішньогрупову дисперсію: $S_W^2 = \frac{1}{N-a} \times$ $\times SSW$, а $S^2 = \frac{1}{N-1} SST$, тоді $S_W^2 = (1-\delta)\, S^2$, або $\frac{S_W^2}{S^2} = 1 - \delta$. Коефіцієнт δ можна використовувати незалежно від того, чи є групи одного розміру, чи ні. При цьому

$$\begin{aligned} &\delta_{\max} = 1 \Rightarrow SSW = 0; \\ &\delta_{\min} = 1 - \tfrac{N-1}{N-a} = -\tfrac{a-1}{N-a} \Rightarrow SSB = 0. \end{aligned}$$

***Твердження* 4.2.** *При вибірковому дизайні* СВ, *коли* $a = \frac{N}{n}$ – *ціле число, дисперсія* π*-оцінки має вигляд*

$$\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right) = \frac{N^2 \cdot S^2}{n} \cdot \left[(1-f) + (n-1) \cdot \delta\right], \quad \textit{де} \quad f = \frac{n}{N} = \frac{1}{a}.$$

Доведення цього твердження залишимо як вправу для самостійного опрацювання.

Порівняємо ефективність систематичного відбору з простим випадковим без повернення:

$$\mathcal{D}_{\text{ПВВбП}}\left(\widehat{t}_\pi\right) = N^2 \frac{1-f}{n} S^2.$$

$$deff\left(\text{СВ},\, \hat{t}_\pi\right) = \frac{\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)}{\mathcal{D}_{\text{ПВВбП}}\left(\widehat{t}_\pi\right)} = 1 + \frac{n-1}{1-f}\delta.$$

Систематичний відбір є ефективнішим за простий випадковий відбір без повернення, якщо $\delta < 0$, тобто, коли $S_W^2 > S^2$.

Систематичний відбір є менш ефективним за ПВВбП, якщо $\delta > 0$, тобто, коли $S_W^2 < S^2$.

Систематичний відбір має таку саму ефективність, як і простий випадковий відбір без повернення, якщо $\delta = 0$, тобто тоді, коли $S_W^2 = S^2$.

4.5. Оцінювання дисперсії при систематичному відборі

Одним із головних недоліків систематичного відбору, що компенсує простоту реалізації відбору, є неможливість оцінити дисперсію. Існує декілька способів вирішення цієї проблеми, але жодний з них не є достатньо хорошим. Наведемо два з них.

Спосіб 1. Зміщені оцінки для дисперсії

Наведемо одну з можливих зміщених оцінок для $\mathcal{D}(\widehat{t}_\pi)$.

Припустимо, що систематичний відбір настільки ефективний, як і простий випадковий, тобто $\delta \approx 0$. Якщо s_r – це вибірка при систематичному відборі, то вибірковою дисперсією є

$$\widehat{S}_r^2 = \frac{1}{n-1} \sum_{k \in s_r} (y_k - \bar{y}_{s_r})^2.$$

Тоді за оцінку дисперсії $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$ можемо вибрати

$$\widehat{\mathcal{D}} = N^2 \frac{1-f}{n} \widehat{S}_r^2.$$

Запропонована оцінка дисперсії має такий самий вигляд, як і у випадку простого випадкового відбору без повернення.

У випадку, коли систематичний відбір є більш ефективним за простий випадковий, тобто, якщо $\delta < 0$, то

$$\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right) < \mathcal{D}_{\text{ПВВбП}}\left(\widehat{t}_\pi\right).$$

Тоді $\widehat{\mathcal{D}}$ буде переоцінювати реальне значення $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$, тобто

$$E_{\text{СВ}}\left(\widehat{\mathcal{D}}\right) > \mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right).$$

У цьому випадку при побудові довірчих інтервалів рівень довіри $1-\alpha$ буде в дійсності більшим, ніж припускається.

У випадку, коли $\delta > 0$, $\widehat{\mathcal{D}}$ буде недооцінювати дійсне значення $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$, і при побудові довірчих інтервалів, рівень довіри $1-\alpha$ буде в дійсності меншим, ніж припускається.

При $\delta = 0$ оцінка $\widehat{\mathcal{D}}$ співвідноситься з $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$ досить добре.

Фактично ефективність систематичного відбору залежить від упорядкованості елементів генеральної сукупності. Якщо елементи в списку впорядковані так, що немає жодної явної залежності між номером елемента і значенням характеристики y, то елементи в систематичній вибірці не матимуть також цієї залежності і властивості такої вибірки будуть мало відрізнятись від вибірки при ПВВбП. Тому в цьому випадку використання оцінки $\widehat{D}$ може бути виправданим.

Якщо елементи в генеральній сукупності впорядковані так, що значення y_k зростають (або спадають), тоді систематичний відбір буде навіть більш ефективним ніж ПВВбП, і оцінка $\widehat{D}$ буде переоцінювати реальне значення дисперсії.

Найгірша ситуація може виникнути тоді, коли у впорядкованих значеннях характеристики y спостерігається деяка періодичність, і крок, з яким робиться відбір (a), дорівнює періоду. Тоді, оцінка $\widehat{D}$ буде недооцінювати реальне значення дисперсії.

Таким чином, оцінкою $\widehat{D}$ рекомендується користуватись лише в тих випадках, коли є впевненість у тому, що елементи генеральної сукупності досить добре перемішані і, що дуже важливо, дані не мають ніякої періодичності.

Спосіб 2. Модифікація систематичного відбору

Інший підхід – це модифікувати систематичний відбір так, щоб можна було скористатись класичною оцінкою $\widehat{\mathcal{D}}\left(\widehat{t}_\pi\right)$.

Замість того, щоб використовувати один випадковий старт r та вибірковий інтервал a, використовують $m > 1$ випадкових стартів та вибірковий інтервал ma. У результаті цього отримуємо m «розшарувань», кожне розміру $\frac{n}{m}$.

Припустимо для простоти, що $\frac{n}{m}$ та $a = \frac{N}{n}$ – цілі числа.

Випадковим чином вибирається m чисел від 1 до ma:

$$r_1, r_2, \ldots, r_m.$$

Тоді вибірка буде формуватись так:

$$s = \left\{k : \; k = r_i + (j-1)\, ma : i = \overline{1,m}, \; j = \overline{1, n/m}\right\}.$$

У цьому випадку ймовірності включення будуть дорівнювати:

$$\pi_k = \frac{m}{ma} = \frac{1}{a} = \frac{n}{N},$$

$$\pi_{kl} = \begin{cases} \frac{n}{N}, & \text{якщо } k \text{ та } l \text{ належать до одного розшарування;} \\ \frac{n}{N}\frac{m-1}{ma-1}, & \text{якщо } k \text{ та } l \text{ належать до різних розшарувань.} \end{cases}$$

При такій модифікації дизайн буде вже вимірним і ми можемо скористатись незміщеною оцінкою для дисперсії Горвіца–Томпсона.

По суті, систематичний відбір є частинним випадком кластерного відбору, який будемо розглядати в розділі 8. Розшарування відіграють роль кластерів. Випадок систематичного відбору з одним випадковим стартом відповідає одностадійному кластерному відбору одного кластера, який повністю обстежується. Коли ми використовуємо декілька випадкових стартів, це відповідає випадковому відбору декількох кластерів. Отже, для оцінювання дисперсії при систематичному відборі з декількома випадковими стартами можна скористатись формулою (8.11) для одностадійного кластерного відбору.

Основним недоліком цього підходу є те, що оцінки при його використанні часто мають більшу дисперсію порівняно з систематичним відбором з одним випадковим стартом.

У деяких випадках проблеми з оцінюванням не є настільки серйозними. Зокрема, коли $\delta \approx 0$.

Часто систематичний відбір використовується на останніх стадіях складного обстеження.

4.6. Вправи та питання для самоконтролю

4.1. Яка основна перевага систематичного відбору?

4.2. Чому при СВ не можна використовувати оцінку для дисперсії π-оцінки?

4.3. В яких випадках варто, а в яких не варто застосовувати зміщену оцінку для дисперсії при використанні систематичного відбору? Чому?

4.4. Для генеральної сукупності розміру $N = 10$ виписати всі можливі систематичні вибірки розміру $n = 4$, що можуть бути отримані при застосуванні:

1) методу дробового вибіркового інтервалу;

2) циклічного методу.

Обчислити для обох методів всі ймовірності включення першого та другого порядків.

4.5. Для обстеження стану книг у бібліотеці з фондом 12 000 книг було застосовано систематичний відбір з двома випадковими стартами. Простим випадковим чином було обрано два числа (випадкових старти) з проміжку від 1 до 30, які визначали реєстраційний номер книги в бібліотеці, а саме – 5 та 26. Кожному з них відповідає одна систематична вибірка: $r_1 = 5$ відповідає $s_1 = \{5, 35, ...\}$ та $r_2 = 26$ відповідає $s_2 = \{26, 56, ...\}$. Для кожної з цих систематичних вибірок було пораховано кількість книг у поганому стані: $N_{s_1} = 8$ та $N_{s_2} = 13$.

Оцінити загальну кількість книг у поганому стані в бібліотеці та обчислити значення оцінки дисперсії.

4.6. Обстежується 280 середньоосвітніх шкіл в деякій області України з метою визначення середнього рівня знань учнів по 12-бальній шкалі. Для цього систематичним відбором з вибірковим інтервалом $a = 10$ відбираються вибірки при різній упорядкованості (уп.) списку шкіл:

уп. 1 – в алфавітному порядку населених пунктів, де вони розміщені; якщо в одному населеному пункті є декілька шкіл, то додатково враховувався номер цієї школи;

уп. 2 – у порядку зростання залежно від кількості учнів, що навчаються в школі;

уп. 3 – у порядку зростання залежно від кількості випускників-медалістів за останні 5 років.

Для кожного такого впорядкування існує 10 можливих систематичних вибірок. У табл. 7 наведено значення середньої успішності в школах кожної систематичної вибірки для кожного з трьох упорядкувань.

Таблиця 7. **Середня успішність шкіл $\overline{y}$ для систематичних вибірок**

Вибірка	уп. 1	уп. 2	уп. 3
1	8,826	8,507	8,653
2	8,938	9,006	8,770
3	8,531	9,503	9,036
4	11,821	9,736	8,873
5	11,051	10,149	9,172
6	9,235	9,673	9,504
7	9,592	9,047	9,222
8	8,432	8,538	9,443
9	7,037	8,499	9,639
10	8,411	9,216	9,562

1) Підрахувати дисперсію оцінки Горвіца–Томпсона при кожному із трьох упорядкувань.

2) Нехай дисперсія генеральної сукупності для середньої успішності $S^2 = 1,392$. Обчислити дисперсію оцінки Горвіца–Томпсона при ПВВбП з $n = 28$.

3) Обчислити коефіцієнт міри однорідності δ, визначений у формулі (4.2), для кожного з трьох упорядкувань. Яким буде мінімально можливе значення цього коефіцієнта?

4.7. Довести твердження 4.2.

Розділ 5
Відбір з поверненням

5.1. Основні відмінності

До цього часу при відборі елементів у вибірку ми не давали жодного шансу елементу потрапити туди більше одного разу (тобто, відбір був без повернення). І це логічно – при повторному включенні елемента ми не залучаємо до вибірки нової інформації. Але відбір з поверненням має право на існування і детальне вивчення. Перш за все тому, що в цьому випадку деякі оцінки мають дуже прості статистичні властивості, що дає можливість досліджувати досить складні процедури відбору.

Дослідимо фундаментальні особливості двох відборів – *без* та *з* поверненням на прикладі простого випадкового відбору.

Приклад 5.1. Розглянемо схему відбору, при якій виконується m незалежних відборів елементів із генеральної сукупності розміру N з однаковими ймовірностями $1/N$. Відібраний елемент повертається в сукупність. Таким чином всі N елементів беруть участь у відборі на кожному кроці. Такий відбір назвемо *простим випадковим відбором з поверненням* (англ. `simple random sampling with replacement`), скорочено ПВВзП.

При ПВВзП кожен елемент має додатну ймовірність бути обраним у вибірку більше ніж один раз. Тоді ймовірність того, що елемент був обраний n разів при m спробах дорівнює

$$C_m^n \left(\frac{1}{N}\right)^n \left(1 - \frac{1}{N}\right)^{m-n}.$$

Ймовірність того, що елемент не був обраний: $(1 - \frac{1}{N})^m$, отже ймовірність того, що елемент був обраний хоча б один раз, дорівнює

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m, \quad k = 1, ..., N.$$

$\Diamond$

При розгляді відбору з поверненням потрібно уточнити, що саме ми розуміємо під словом «вибірка». Якщо k_i – елемент, відібраний на i-му кроці, то $os = (k_1, k_2, ..., k_m)$ – це впорядкована вибірка (англ. `ordered sample`). Вона несе інформацію про порядок відбору елементів, а також про повтори. Ймовірнісний розподіл упорядкованих вибірок називається *впорядкованим вибірковим дизайном*.

Оскільки при повторному включенні елемента у вибірку не отримується додаткової інформації, тому можна розглядати вибірку без повторень та порядку $s = \{k : k = k_i, i = \overline{1,m}\}$. Розмір такої вибірки n_s є випадковою величиною, причому $n_s \le m$ з імовірністю 1.

У деяких випадках оцінки мають дуже прості статистичні властивості для впорядкованих вибірок os, у той час як для невпорядкованих вибірок s досліджувати ці властивості складно. Проілюструємо це на прикладі ПВВзП.

Приклад 5.2. При ПВВзП з m випробувань можливо утворити N^m різних, але рівноймовірних упорядкованих вибірок розміру m. Отже,

$$p(os) = \begin{cases} \frac{1}{N^m}, & \text{для всіх упорядкованих вибірок } os \text{ розміру } m; \\ 0, & \text{в іншому випадку.} \end{cases}$$

Таким чином, упорядкований вибірковий дизайн, що відповідає простому випадковому відбору з поверненням – це рівномірний розподіл на множині всіх можливих упорядкованих вибірок розміру m.

На противагу цьому, розподіл неупорядкованої вибірки s, що виникає при такому відборі має набагато складніший розподіл. Але для практичного використання нам не потрібно знати безпосередньо вибірковий дизайн. Достатньо підрахувати ймовірності включення першого та другого порядку

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m, \quad k = 1, ..., N,$$

та

$$\pi_{kl} = 1 - 2\left(1 - \frac{1}{N}\right)^m + \left(1 - \frac{2}{N}\right)^m, \quad k \ne l = 1, ..., N.$$

◊

Для відбору з поверненням дуже просто перейти до відбору, що допускає нерівні ймовірності відбору елементів, зберігаючи незалежність відборів на кожному кроці.

Нехай $p_1, p_2, ..., p_N$ – набір наперед заданих додатних чисел, що задовольняють умову $\sum_{k \in U} p_k = 1$. Тоді процедура відбору проводиться так, що на першому кроці

$$P\{\text{відібраний } k\text{-й елемент}\} = p_k, \quad k = \overline{1,N}.$$

Відібраний таким чином елемент k_1 повертається у сукупність. Той самий набір ймовірностей використовується на кожному наступному кроці відбору. Отже, ймовірність отримати фіксовану впорядковану вибірку $(k_1, k_2, ..., k_m)$

$$P\{(k_1, k_2, ..., k_m)\} = p_{k_1} p_{k_2} ... p_{k_m}.$$

Звідси досить просто отримати вибірковий дизайн для впорядкованих вибірок os. Для вибірок, що не враховують повтори та порядок, такий розподіл є дуже складним. Обчислимо ймовірність включення елемента у вибірку, тобто, ймовірність того, що елемент потрапить у вибірку хоча б один раз:

$$\pi_k = 1 - (1 - p_k)^m.$$

Якщо $m = 1$, то $\pi_k = p_k$. Якщо $m > 1$, а значення p_k – досить малі, то $(1 - p_k)^m \approx 1 - mp_k$, отже, $\pi_k \approx mp_k$.

5.2. Оцінка Хансена–Гурвіца

Для того, щоб оцінити сумарне $T = \sum_{k\in U} y_k$ можна скористатись оцінкою Горвіца–Томпсона. Але при $m > 1$ підрахунок ймовірностей π_k є досить трудомісткою справою. Можна побудувати дещо іншу оцінку.

Розглянемо p-зважене значення k-го елемента (англ. p-expended value)

$$\frac{y_k}{p_k}.$$

Тобто, на кожен елемент «навішуються» ваги, що обернено пропорційні ймовірності його включення у вибірку на кожному кроці, щоб компенсувати те, що не всі елементи обстежуються. Усереднивши ці значення, отримаємо нову оцінку

$$\widehat{t}_{pwr} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}}, \tag{5.1}$$

що має назву оцінки Хансена–Гурвіца[3]. Зауважимо, що при $m > 1$ $\pi_k \neq p_k$ та сума в (5.1) береться по всіх елементах, що потрапили у впорядковану вибірку, незважаючи на повтори.

***Твердження* 5.1.** *Оцінка Хансена–Гурвіца*

$$\widehat{t}_{pwr} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}} \tag{5.2}$$

є незміщеною оцінкою сумарного T та має дисперсію

$$\mathcal{D}(\widehat{t}_{pwr}) = \frac{1}{m}V_1, \text{ де } \quad V_1 = \sum_{k\in U}\left(\frac{y_k}{p_k} - T\right)^2 p_k, \tag{5.3}$$

яку можна незміщено оцінити

$$\widehat{\mathcal{D}}(\widehat{t}_{pwr}) = \frac{1}{m}\widehat{V}_1, \text{ де } \quad \widehat{V}_1 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_{ki}}{p_{ki}} - \widehat{t}_{pwr}\right)^2. \tag{5.4}$$

[3]Скорочення *pwr* походить від англійського виразу p-expended with replacement

Доведення. Нехай випадкові величини $Z_i = \frac{y_k}{p_k}$, якщо k-й елемент вибраний на i-му кроці. Тоді для кожного $i = \overline{1,m}$

$$P\left(Z_i = \frac{y_k}{p_k}\right) = p_k, \quad k = \overline{1,N}.$$

Випадкові величини Z_i, $i = \overline{1,N}$ незалежні, однаково розподілені та

$$\begin{aligned} EZ_i &= \sum_{k\in U}\frac{y_k}{p_k}p_k = T; \\ \mathcal{D}Z_i &= E(Z_i - T)^2 = \sum_{k\in U}\left(\frac{y_k}{p_k} - T\right)^2 p_k = V_1. \end{aligned}$$

Оскільки $\widehat{t}_{pwr} = \frac{1}{m}\sum\limits_{i=1}^{m} Z_i = \overline{Z}$ – це арифметичне середнє, то

$$\begin{aligned} E(\widehat{t}_{pwr}) &= \frac{1}{m}\sum_{i=1}^{m} EZ_i = \frac{1}{m}mT = T; \\ \mathcal{D}(\widehat{t}_{pwr}) &= \frac{1}{m^2}\sum_{i=1}^{m}\mathcal{D}Z_i = \frac{1}{m^2}mV_1 = \frac{V_1}{m}. \end{aligned}$$

Для доведення незміщеності $\widehat{\mathcal{D}}(\widehat{t}_{pwr})$ достатньо довести, що $E\widehat{V}_1 = V_1$:

$$\begin{aligned} \widehat{V}_1 &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}} - \widehat{t}_{pwr}\right)^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Z_i - \overline{Z})^2 = \\ &= \frac{1}{m-1}\sum_{i=1}^{m}(Z_i - T - (\overline{Z} - T))^2 = \\ &= \frac{1}{m-1}\left[\sum_{i=1}^{m}(Z_i - T)^2 - m(\overline{Z} - T)^2\right]. \end{aligned}$$

Тоді

$$E\widehat{V}_1 = \frac{1}{m-1}\left[\sum_{i=1}^{m} E\left(Z_i - T\right)^2 - mE(\overline{Z} - T)^2\right] =$$
$$= \frac{1}{m-1}\left[\sum_{i=1}^{m} \mathcal{D}Z_i - m\mathcal{D}\overline{Z}\right] = \frac{1}{m-1}\left[mV_1 - m\frac{1}{m}V_1\right] = V_1.$$

□

Зауваження 5.1. Альтернативною оцінкою для відбору з поверненням є оцінка Горвіца–Томпсона. Для $m \geq 2$

$$\widehat{t}_{\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} \neq \widehat{t}_{pwr}.$$

Яка з них краща? Простої відповіді на це запитання немає. Вони обидві незміщені, а дисперсія цих оцінок залежить від конкретного набору значень y_k.

5.3. Оцінка Хансена–Гурвіца при простому випадковому відборі з поверненням

Сформулюємо тепер основні результати для випадку, коли будь-яка впорядкована вибірка $os = (k_1, k_2, \ldots, k_m)$ має однакову ймовірність бути обраною.

***Твердження* 5.2.** *При простому випадковому відборі з поверненням оцінка Хансена–Гурвіца для сумарного T має вигляд*

$$\widehat{t}_{pwr} = N \cdot \overline{y}_{os},$$

де $\overline{y}_{os} = \frac{1}{m}\sum_{i=1}^{m} y_{k_i}$ – середнє впорядкованої вибірки.

Дисперсія та її оцінка при цьому мають вигляд:

$$\mathcal{D}_{\text{ПВВзП}}\left(\widehat{t}_{pwr}\right) = N\left(N-1\right)\frac{S^2}{m};$$

$$\widehat{\mathcal{D}}_{\text{ПВВзП}}\left(\widehat{t}_{pwr}\right) = N^2\frac{\widehat{S}^2_{os}}{m},$$

де S^2 – дисперсія генеральної сукупності,
$\widehat{S}^2_{os} = \frac{1}{m-1}\sum_{i=1}^{m}\left(y_{k_i} - \overline{y}_{os}\right)^2$ – дисперсія впорядкованої вибірки.

Доведення. При ПВВзП $p_k = \frac{1}{N}$. Тоді величини V_1 та $\widehat{V}_1$ у формулах (5.3) та (5.4) набувають вигляду

$$V_1 = \sum_{k \in U}\left(\frac{y_k}{p_k} - T\right)^2 p_k = \frac{1}{N}\sum_{k \in U}\left(Ny_k - N\overline{Y}\right)^2 =$$
$$= N\sum_{k \in U}\left(y_k - \overline{Y}\right)^2 = N\left(N-1\right)S^2;$$
$$\widehat{V}_1 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Ny_{k_i} - N\overline{y}_{os}\right)^2 = N^2\widehat{S}^2_{os}.$$

Звідси і випливає твердження. □

Порівняємо методи простого випадкового відбору *з* та *без* повернення. Нехай розмір вибірки n при відборі без повернення дорівнює кількості елементів m у впорядкованій вибірці os. Тоді

$$\frac{\mathcal{D}_{\text{ПВВзП}}\left(\widehat{t}_{pwr}\right)}{\mathcal{D}_{\text{ПВВбП}}\left(\widehat{t}_{\pi}\right)} = \frac{N\left(N-1\right)\frac{S^2}{n}}{N^2\left(1-f\right)\frac{S^2}{n}} =$$
$$= \frac{N-1}{N\left(1-f\right)} = \frac{1-\frac{1}{N}}{1-f} \approx \frac{1}{1-f}.$$

Якщо розмір генеральної сукупності N досить великий, а частка відбору $f = \frac{n}{N}$ досить мала, то ці дві вибіркові стратегії мають майже однакову ефективність. Якщо ж f досить значне, то при відборі з поверненням ефективність втрачається. Наприклад, при $f = 0,5$ дисперсія при ПВВзП удвічі більша, ніж при ПВВбП.

Приклад 5.3. Розглянемо сукупність із прикладу 2.1 та застосуємо простий випадковий відбір з поверненням. У цьому випадку отримаємо вже $6 \times 6 = 36$ можливих упорядкованих вибірок.

Табл. 8 складена таким чином, що вибірки 1–15 та 16–30 складаються з одних і тих самих елементів, але впорядкування у них різне. Ми розрізняємо ці вибірки, але по суті вони ідентичні. Новими є вибірки 31–36. Вони складаються з однакових елементів.

Таблиця 8. **36 можливих вибірок з поверненням розміру 2 із сукупності розміру 6**

Вибірка	1	2	3	4	5	6	7	8	9	10	11	12
Значення	2	2	2	2	2	6	6	6	6	8	8	8
	6	8	10	10	12	8	10	10	12	10	10	12
Вибірка	13	14	15	16	17	18	19	20	21	22	23	24
Значення	10	10	10	6	8	10	10	12	8	10	10	12
	10	12	12	2	2	2	2	2	6	6	6	6
Вибірка	25	26	27	28	29	30	31	32	33	34	35	36
Значення	10	10	12	10	12	12	2	6	8	10	10	12
	8	8	8	10	10	10	2	6	8	10	10	12

Дослідимо, який вигляд матиме вибірковий розподіл оцінки середнього при такому вибірковому дизайні (табл. 9). Для цього скористаємося оцінкою Хансена–Гурвіца:

$$\widehat{\overline{y}}_{pwr} = \frac{1}{N}\widehat{t}_{pwr} = \frac{1}{N}\frac{1}{m}\sum_{i=1}^{m}\frac{y_k}{1/N} = \frac{1}{m}\sum_{i=1}^{m} y_k = \overline{y}_{os}.$$

Таблиця 9. **Вибірковий розподіл $\widehat{\overline{y}}_{pwr}$**

$\widehat{\overline{y}}_{pwr}$	2	4	5	6	7	8	9	10	11	12
P	1/36	2/36	2/36	5/36	4/36	5/36	6/36	6/36	4/36	1/36

Усереднивши, отримаємо дійсне значення параметра $\overline{Y} = 8$ в силу незміщеності оцінки Хансена–Гурвіца. Але, як видно з розподілу цієї оцінки, розкид можливих значень порівняно з відбором без повернення буде більшим. Зокрема, збільшиться кількість «поганих» вибірок. А також це відобразиться на значенні дисперсії оцінки. У цьому випадку $\mathcal{D}(\widehat{\overline{y}}_{pwr}) = \frac{16}{5} \approx 5{,}3$, в той час як при відборі без повернення $\mathcal{D}(\widehat{\overline{y}}_{\pi}) = \frac{64}{15} \approx 4{,}3$. ◊

Отже, все наведене вище говорить не на користь такого відбору. Чому ж вивчення відбору з поверненням все ж таки є важливим?

По-перше, різниця в точності оцінок при відборі *з* та *без* повернення при великих розмірах генеральної сукупності та вибірки не настільки суттєва. А спрощення математичних викладок для відбору з поверненням є досить великою перевагою, враховуючи той факт, що зазвичай для великих сукупностей застосовуються складні методи оцінювання.

По-друге, при роботі з малими сукупностями виникають проблеми з побудовою довірчих інтервалів. Центральна гранична теорема, що зазвичай використовується для цього, не може бути застосована оскільки вона задає граничний розподіл при досить великому розмірі вибірки n. При відборі *з* поверненням ми можемо отримати вибірки будь-якого розміру, в той час як при відборі *без* повернення розмір вибірки не може бути більшим за розмір генеральної сукупності. І взагалі, при відборі з поверненням немає ніякої різниці між, наприклад, сукупностями

$$U_1 = \{0,1,1\},\ U_2 = \{0,0,1,1,1,1\} \text{ та } U_3 = \{0,0,0,1,1,1,1,1,1\}.$$

А отже, ми можемо перейти до сукупностей нескінченного розміру. І в цьому випадку *відбір з поверненням буде еквівалентний відбору без повернення із сукупності нескінченного розміру.*

5.4. Відбір, p-пропорційний до розміру

При простому випадковому відборі з поверненням кожен елемент генеральної сукупності має однакову ймовірність потрапити у впорядковану вибірку на кожному кроці відбору. Але оцінка Хансена–Гурвіца може бути застосована і тоді, коли елементи мають різні ймовірності p_k бути відібраними у вибірку. Тоді постає питання стосовно того, яким чином краще вибрати ці ймовірності, для того щоб ефективність (точність) оцінки Хансена–Гурвіца була найвищою.

$\mathcal{D}(\widehat{t}_{pwr}) = 0$, якщо $y_k = cp_k$ для всіх $k = 1, ..., N$, де c – деяка константа. Тобто, найточнішою буде оцінка $\widehat{t}_{pwr}$, коли p_k пропорційні y_k.

Оскільки y_k не відомі до обстеження, тому на практиці p_k вибирають так, щоб вони були приблизно пропорційні y_k. Оцінка Хансена–Гурвіца $\widehat{t}_{pwr}$ має свої переваги в тому випадку, коли відомі значення деякої допоміжної змінної x: $x_1, x_2, ..., x_N$ такі, що $\frac{y_k}{x_k} \approx \text{const}$. Тоді ймовірності p_k вибираються так, щоб $p_k = cx_k$. Оскільки

$$1 = \sum_{k\in U} p_k = c\sum_{k\in U} x_k, \quad \text{то} \quad c = \frac{1}{\sum\limits_{k\in U} x_k}.$$

Отже,

$$p_k = \frac{x_k}{\sum\limits_{i\in U} x_i}, \quad k \in U. \tag{5.5}$$

Ймовірності p_k, визначені формулою (5.5), називаються *ймовірностями, пропорційними до розміру* (англ. `probability proportional-to-size`), оскільки найчастіше характеристика x є деякою мірою розміру k-го елемента.

У випадку, коли для оцінки Хансена–Гурвіца використовують ймовірності (5.5), такий відбір називається *p-пропорційним до розміру* (англ. `p-proportional-to-size sampling`, або скорочено *pps*).

Приклад 5.4.

1) При обстеженні доходів підприємств (y) допоміжною мірою розміру x можуть бути загальні активи або кількість працівників на підприємстві;

2) при обстеженні врожаю (y), x – це площа земель, що обробляються;

3) оборот магазинів (y) залежить від торгової площі магазину x, і т. д. ◊

***Твердження* 5.3.** *При застосуванні відбору, p-пропорційного до розміру, оцінка Хансена–Гурвіца для сумарного T буде мати вигляд*

$$\widehat{t}_{pwr} = \frac{\sum_{j\in U} x_j}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{x_{k_i}}$$

Зауваження 5.2. Звертаємо увагу читача на те, що при використанні оцінки Горвіца–Томпсона та відбору без повернення ймовірності включення π_k теж можна вибрати пропорційними допоміжній змінній x. Такий відбір називається *π-пропорційним до розміру* (англ. `π-proportional-to-size sampling`, або скорочено *πps*).

Наступне питання, що виникає при використанні різних ймовірностей для відбору: за допомогою якої схеми відбору можна отримати вибірку елементів, що потрапляють у неї із заданими ймовірностями p_k? Наведемо два методи, що відповідають цій вимозі – *метод накопичених сум* та *метод Лахірі*.

Метод накопичених сум. Нехай $T_0 = 0$, $T_k = T_{k-1} + x_k$, $k = 1, 2, ..., N$. Генеруємо рівномірно розподілене на [0,1] випадкове число ε. Якщо $T_{k-1} < \varepsilon T_N \le T_k$, то елемент k потрапляє у вибірку *os* і повертається в генеральну сукупність. Процедура повторюється m разів. При цьому ймовірність того, що елемент k потрапить у вибірку при одному випробуванні, така:

$$p_k = P\{T_{k-1} < \varepsilon T_N \le T_k\} = \frac{T_k - T_{k-1}}{T_N} = \frac{x_k}{\sum\limits_{i\in U} x_i}.$$

Приклад 5.5. Нехай генеральна сукупність складається з п'яти елементів, для кожного з яких відоме значення допоміжної змінної x. Підрахуємо накопичені суми та відповідні інтервали (табл. 10).

Таблиця 10. **Значення допоміжної змінної x_k, накопичені суми T_k та відповідні інтервали**

k	x_k	T_k	Інтервал
1	5	5	[0 ; 5]
2	7	12	(5 ; 12]
3	10	22	(12 ; 22]
4	3	25	(23 ; 25]
5	9	34	(26 ; 34]

Припустимо, що при першому випробуванні ми отримали випадкове число $\varepsilon_1 = 0,214$. Тоді значення $\varepsilon_1 T_5 = 0,214 \cdot 34 = 7,274$ потрапляє в інтервал, що відповідає елементу 2. Отже, на першому кроці елемент 2 потрапляє у вибірку, і так далі, поки не отримаємо необхідну кількість елементів у вибірці. ◊

Якщо N є досить великим, то застосування цього методу відбору без використання комп'ютера може бути досить трудомістким. У цьому випадку перевагу найчастіше віддають методу, запропонованому Лахірі.

Метод Лахірі. Нехай число M таке, що $M \geq \max(x_1, x_2, ..., x_N)$. Відбір одного елемента проводиться у два кроки.

Крок 1. Генеруємо випадкове число j від 1 до N;
Крок 2. Генеруємо випадкове число R від 1 до M.

Якщо $R \leq x_j$, то j-й елемент потрапляє у вибірку. Якщо ж $R > x_j$, то кроки 1 і 2 повторюються.

Означення 5.1. Випробування при застосуванні методу Лахірі будемо називати ефективним, якщо при цьому у вибірку потрапляє один елемент.

Теорема 5.1. *Ймовірність вибрати k-й елемент при першому ефективному випробуванні за методом Лахірі така:*

$$p_k = \frac{x_k}{\sum\limits_{i \in U} x_i}.$$

Доведення. Випробування стає неефективним, якщо на першому кроці ми отримуємо елемент k, а на другому кроці – $R > x_k$. Тоді ймовірність того, що випробування буде неефективним:

$$\lambda = \sum_{k=1}^{N} \frac{1}{N} \left(\frac{M - x_k}{M} \right) = 1 - \frac{\bar{X}}{M},$$

де $\bar{X} = \frac{1}{N} \sum\limits_{i \in U} x_i$ – середнє по змінній x. Звідси випливає, що ймовірність вибору k-го елемента при першому ефективному випробуванні

$$\frac{x_k}{NM} + \frac{x_k}{NM}\lambda + \frac{x_k}{NM}\lambda^2 + ... = \frac{x_k}{NM}(1 - \lambda)^{-1} =$$
$$= \frac{x_k}{NM}\left(1 - 1 + \frac{\bar{X}}{M}\right)^{-1} = \frac{x_k}{\sum\limits_{i \in U} x_i}.$$

□

Цю процедуру повторюють доти, поки у вибірці не буде m елементів. При кожному ефективному випробуванні вибраний елемент повертається в генеральну сукупність, оскільки ми проводимо відбір з поверненням.

5.5. Вправи та питання для самоконтролю

5.1. Чим відрізняється впорядкований вибірковий дизайн від просто вибіркового дизайну?

5.2. Якого вигляду набуває оцінка Горвіца–Томпсона при ПВВзП?

5.3. Чи може оцінка Хансена–Гурвіца бути використаною при відборі з поверненням?

5.4. Розглянемо випадок простого випадкового відбору з поверненням з m відборами з генеральної сукупності розміру N. Довести, що $P\{k \in os\} = \frac{m}{N} + O\left(\frac{m^2}{N^2}\right)$. Проаналізувати випадок, коли N набуває великих значень.

5.5. Нехай генеральна сукупність складається з N елементів. Елементи потрапляють у вибірку за допомогою простого випадкового відбору з поверненням з $m = 3$ елементами в упорядкованій вибірці os.

Розглянемо функцію $r(\cdot)$, яка вилучає з упорядкованої вибірки інформацію про порядок і повтори. Наприклад:

$$r(\{2,2,3\}) = \{2,3\}, \quad r(\{3,2,3\}) = \{2,3\}, \quad r(\{3,3,3\}) = \{3\}.$$

Після дії функцією r на впорядковану вибірку os ми отримуємо невпорядковану вибірку без повторів – s.

1) Підрахувати ймовірність R_i, того, що вибірка os буде містити i різних елементів, $i = 1, 2, 3$.

2) Довести, що умовний вибірковий дизайн вибірок s за умови фіксованого розміру вибірки n_s – це ПВВбП.

3) Виписати вибірковий дизайн для s.

4) Розглянути дві оцінки для середнього: $\overline{y}_{pwr} = \frac{1}{3}\sum_{k\in os} y_k$ – середнє з повторами; $\overline{y}_\pi = \frac{1}{n_s}\sum_{k\in s} y_k$ – середнє по різних елементах із вибірки. Порахувати математичне сподівання та дисперсію цих оцінок. Який висновок можна зробити?

5.6. При ПВВзП для оцінювання сумарного T можна скористатись двома незміщеними оцінками

$$\widehat{t}_1 = \frac{N}{m}\sum_{i=1}^{m} y_{k_i} \quad \text{та} \quad \widehat{t}_2 = \frac{\sum_{k\in s} y_k}{1-\left(1-\frac{1}{N}\right)^m}.$$

Для цих оцінок не можна стверджувати, що для всіх векторів $\mathbf{y} = (y_1, y_2, \ldots, y_N) \in \mathbb{R}^N$ дисперсія однієї оцінки менша (чи більша), ніж дисперсія іншої. Навести приклади таких векторів $\mathbf{y}'$ та $\mathbf{y}''$ з $\mathbb{R}^N$, що

1) $\mathcal{D}_{\text{ПВВзП}}(\widehat{t}_1) < \mathcal{D}_{\text{ПВВзП}}(\widehat{t}_2)$ для $\mathbf{y}'$;

2) $\mathcal{D}_{\text{ПВВзП}}(\widehat{t}_1) > \mathcal{D}_{\text{ПВВзП}}(\widehat{t}_2)$ для $\mathbf{y}''$.

Розділ 6

Методи нерівноймовірнісного відбору без повернення

Відбір Бернуллі, простий випадковий відбір *з* та *без* повернення, систематичний відбір – всі ці методи відбору використовують однакові ймовірності включення першого порядку $\pi_k = \text{const}$, $k \in U$. Це зумовлює дуже простий вигляд оцінок. Але ця властивість є далеко не обов'язковою умовою для відбору. Навпаки, на практиці частіше використовуються методи відбору з неоднаковими ймовірностями, оскільки саме вони є більш ефективними.

При розгляді відбору без повернення та побудові оцінки Хансена–Гурвіца (див. розділ 5.4.) ми вже зустрічались із випадком, коли ймовірності p_k потрапляння елементів у вибірку були різними. При оптимальному виборі цих ймовірностей можна отримати оцінку з набагато меншою дисперсією, ніж при відборі з однаковими ймовірностями.

У цьому розділі розглянемо методи нерівноймовірнісного відбору, за яких елементи не мають жодного шансу потрапити у вибірку більше одного разу, а саме: відбір Пуассона та відбір, π-пропорційний до розміру. Для оцінювання відповідних параметрів будемо використовувати оцінку Горвіца–Томпсона.

6.1. Відбір Пуассона

Цей метод відбору є узагальненням відбору Бернуллі. Його можна отримати таким чином: кожному елементу $k \in U$ ставиться у відповідність число π_k, $0 \le \pi_k \le 1$, що визначене наперед; при покроковому переборі кожного елемента генеральної сукупності елемент k потрапляє у вибірку з імовірністю π_k та не потрапляє у вибірку з імовірністю $1 - \pi_k$. Тобто,

$$\forall k \quad P\{I_k = 1\} = \pi_k, \quad P\{I_k = 0\} = 1 - \pi_k.$$

Тоді вибірковий дизайн відбору Пуассона (ВП) має вигляд

$$p(s) = \prod_{k\in s} \pi_k \prod_{k\in U\setminus s} (1-\pi_k), \quad s \in \mathfrak{F},$$

де $\mathfrak{F}$ – множина всіх 2^N підмножин генеральної сукупності U. Оскільки випробування незалежні, то $\pi_{kl} = \pi_k\pi_l$ для всіх $k, l \in U$.

При заданих наперед ймовірностях $\pi_1, ..., \pi_N$ відбір Пуассона реалізується за допомогою такої схеми послідовного відбору: генерується N незалежних однаково розподілених на $[0, 1]$ випадкових величин $\varepsilon_1, ..., \varepsilon_N$, що ставляться у відповідність кожному елементу генеральної сукупності. Якщо $\varepsilon_k < \pi_k$, то елемент k включається у вибірку. Якщо ж $\varepsilon_k \geq \pi_k$, то елемент k у вибірку не потрапляє.

При відборі Пуассона, як і при відборі Бернуллі, розмір вибірки є випадковою величиною. При цьому

$$\begin{aligned} En_s &= \sum_{k\in U} \pi_k, \\ \mathcal{D}n_s &= \sum_{k\in U} \pi_k(1-\pi_k). \end{aligned}$$

***Твердження* 6.1.** *При відборі Пуассона оцінка Горвіца–Томпсона для сумарного T має вигляд*

$$\widehat{t}_\pi = \sum_{k\in s} \frac{y_k}{\pi_k}$$

з дисперсією

$$\mathcal{D}_{\text{ВП}}(\widehat{t}_\pi) = \sum_{k\in U} \pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 = \sum_{k\in U}\left(\frac{1}{\pi_k} - 1\right) y_k^2.$$

Незміщенною оцінкою цієї дисперсії є статистика

$$\widehat{\mathcal{D}}_{\text{ВП}}(\widehat{t}_\pi) = \sum_{k\in s} \frac{\pi_k(1-\pi_k)}{\pi_k}\left(\frac{y_k}{\pi_k}\right)^2 = \sum_{k\in s} \frac{1-\pi_k}{\pi_k^2} y_k^2.$$

Зауважимо, що відбір Бернуллі повністю визначений, якщо зафіксувати середній розмір вибірки $En_s = n$ у припущенні, що розмір генеральної сукупності N відомий.

При відборі Пуассона цього замало: при фіксованому середньому розмірі вибірки $En_s = n$ існує безліч варіантів вибору π_k. Який же набір імовірностей $\pi_1, ..., \pi_N$ буде найкращим?

Для визначення цього потрібно розв'язати задачу оптимізації $\mathcal{D}_{\text{ВП}}(\widehat{t}_\pi) \to \min$ при фіксованому середньому розмірі вибірки

$$En_s = n = \sum_{k\in U} \pi_k.$$

Це еквівалентно такій задачі:

$$\left(\sum_{k\in U} \frac{y_k^2}{\pi_k}\right)\left(\sum_{k\in U} \pi_k\right) \to \min.$$

Із нерівності Коші маємо

$$\left(\sum_{k\in U} \frac{y_k^2}{\pi_k}\right)\left(\sum_{k\in U} \pi_k\right) \geq \left(\sum_{k\in U} y_k\right)^2.$$

Знак «=» досягається лише в тому випадку, коли $\frac{y_k}{\pi_k} = \lambda = \text{const}$. Якщо характеристика y набуває лише додатних значень, то $\pi_k = \frac{y_k}{\lambda}$. Отже, з рівностей

$$n = \sum_{k\in U} \pi_k = \frac{1}{\lambda}\sum_{k\in U} y_k, \quad \lambda = \frac{1}{n}\sum_{k\in U} y_k$$

випливає, що

$$\pi_k = \frac{y_k}{\frac{1}{n}\sum_{l\in U} y_l},$$

в припущенні, що

$$y_k \leq \frac{1}{n}\sum_{l\in U} y_l \quad \forall k \in U.$$

Оскільки значення y_k невідомі для всіх k, то цей результат має в основному теоретичну цінність. Тим не менш, якщо відома

деяка додаткова інформація (характеристика x), що досить добре корелює з y, то π_k можна вибрати пропорційно до змінної x:

$$\pi_k = \frac{x_k}{\frac{1}{n}\sum_{l\in U} x_l}, \quad k \in U, \tag{6.1}$$

припускаючи, що

$$x_k \le \frac{1}{n}\sum_{l\in U} x_l,$$

в іншому випадку $\pi_k = 1$.

Якщо $\frac{x_k}{y_k} \approx \text{const}$, то оцінка $\widehat{t}_\pi$ буде мати приблизно мінімальну дисперсію.

Ймовірності включення, визначені формулою (6.1), є *ймовірностями, пропорційними до розміру* (англ. `probability proportional-to-size`), .

Основним недоліком відбору Пуассона, як і відбору Бернуллі, є змінний розмір вибірки. Якщо припустити, що можливо знайти ймовірності π_k оптимальним чином, то

$$\widehat{t}_\pi = \sum_{k\in s}\frac{y_k}{\pi_k} = \sum_{k\in s}\frac{y_k}{y_k/\frac{1}{n}\sum_{l\in U} y_l} = \frac{n_s}{n}\sum_{l\in U} y_l = \frac{n_s}{n}T.$$

Звідси видно, що варіація оцінки $\widehat{t}_\pi$ буде виникати в результаті варіації розміру вибірки.

Це наштовхує на думку, що оцінка $\widehat{t}_\pi$ з вибраними таким чином ймовірностями буде себе добре поводити, при відборах із фіксованим розміром вибірки n_s.

6.2. Відбір, π-пропорційний до розміру

При відборі, π-пропорційному до розміру, ймовірності включення π_k вибираються пропорційно до деякої допоміжної змінної x, яка набуває додатних значень $x_1, x_2, ..., x_N$. Зазвичай змінна x характеризує якусь міру розміру для кожного з елементів генеральної сукупності та корелює з характеристикою y, що вивчається.

При побудові схеми відбору, π-пропорційного до розміру, з фіксованим розміром вибірки, та використанні оцінки Горвіца–Томпсона бажано, щоб виконувались такі умови:

1) безпосередній відбір був порівняно простою процедурою;
2) ймовірності включення першого порядку π_k були строго пропорційними до x_k для всіх $k \in U$;
3) ймовірності включення другого порядку π_{kl} були додатними для всіх $k \neq l$ (вимірність дизайну);
4) всі π_{kl} можна було б підрахувати точно і щоб ці обчислення були не надто складними;
5) $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l < 0$ для всіх $k \neq l$ (умова Єйтса–Гранді–Сена).

У випадку, коли розмір вибірки $n = 1$, імовірності включення першого порядку π_k будуть співпадати з p_k, тобто p-пропорційний до розміру відбір та π-пропорційний до розміру відбір будуть ідентичними. Тому методи накопичених сум та Лахірі можна використати і для відбору, π-пропорційного до розміру.

Для розміру вибірки $n = 2$ існує безліч вибіркових схем для для відбору, π-пропорціного до розміру. Деякі з них досить прості, деякі складні. Ми наведемо лише одну, що була запропонована Брюером. Це схема на основі жеребкування. На кожному кроці ймовірності отримання елемента у вибірку підбираються так, щоб у результаті двох випробувань отримувались бажані ймовірності включення першого порядку:

$$\pi_k = \frac{x_k}{\frac{1}{2}\sum_{i\in U} x_i}.$$

Для простоти будемо вважати, що $x_k < \frac{1}{2}\sum_{i\in U} x_i$ для всіх k.

Схема Брюера [12]. Спочатку для всіх елементів k підраховуються величини

$$c_k = \frac{x_k(T_N - x_k)}{T_N(T_N - 2x_k)}, \text{ де } T_N = \sum_{k\in U} x_k.$$

Крок 1. Елемент k потрапляє у вибірку з імовірністю

$$p_k = \frac{c_k}{\sum_{i\in U} c_i}.$$

Крок 2. Елемент l потрапляє у вибірку за умови того, що на першому кроці був обраний елемент k (без повернення), з імовірністю

$$p_{l|k} = \frac{x_l}{T_N - x_k}.$$

У результаті отримуємо

$$\begin{aligned}\pi_k &= \frac{2x_k}{T_N}, \quad k \in U;\\ \pi_{kl} &= \frac{2x_k x_l (T_N - x_k - x_l)}{T_N \sum_{i\in U} c_i (T_N - 2x_k)(T_N - 2x_l)}, \quad k \neq l.\end{aligned}$$

Дана схема відбору задовольняє всі бажані властивості 1)–5).

Приклад 6.1. Розглянемо генеральну сукупність із прикладу 5.5 та підрахуємо для всіх елементів значення c_k:

k	1	2	3	4	5
x_k	5	7	10	3	9
c_k	0,178	0,278	0,504	0,098	0,413

Нехай у результаті реалізації схеми Брюера у вибірку потрапили елементи 2 та 5. Значення характеристики y для цих елементів – $y_2 = 4$, $y_5 = 6$. Обчислимо значення оцінки Горвіца–Томпсона для сумарного T та оцінку дисперсії з урахуванням того, що:

$$\pi_2 = \frac{14}{34};\ \pi_5 = \frac{18}{34};\ \pi_{25} = \frac{2\cdot 7\cdot 9(34-7-9)}{34\cdot 1,471(34-14)(34-18)} = 0,142;$$

$$\begin{aligned}\widehat{t}_\pi &= \frac{4}{14/34} + \frac{6}{18/34} = 19,8;\\ \frac{\Delta_{25}}{\pi_{25}} &= 1 - \frac{\pi_2\pi_5}{\pi_{25}} = 1 - \frac{\frac{14}{34}\cdot\frac{18}{34}}{0,142} = -0,54.\end{aligned}$$

Використовуючи для оцінки дисперсії форму Єйтса–Гранді–Сена, отримаємо

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = -\frac{1}{2}\cdot 2(-0,54)(8,5-11,3)^2 = 4,23.$$

Як бачимо, навіть при такому малому розмірі вибірки, як $n = 2$, обчислення при оцінюванні є досить трудомісткою справою. ◊

Більшість існуючих схем відбору, π-пропорціного до розміру, з імовірностями, що строго пропорційні до x_k, при $n > 2$ є досить складними. В основному це схеми, в яких використовується жеребкування, і обчислення ймовірностей включення другого порядку дуже ускладнюється при зростанні розміру вибірки.

Наведемо схему, запропоновану Сантером, яка широко використовується на практиці, але в ній вимога строгої пропорційності до x_k послаблена. У цій схемі елементам з основної (важливішої) частини генеральної сукупності надаються ймовірності π_k, що є строго пропорційними до x_k, а всім іншим елементам – однакові ймовірності.

Схема Сантера [21, 22]. Розглянемо генеральну сукупність, що складається з N елементів.

1) Елементи сукупності впорядковуються в порядку спадання відносно значень змінної x. Нехай $\{1, 2, ..., N\}$ – індекси елементів, що вже були впорядковані таким чином.

2) Для елемента $k = 1$ генерується значення ε_1 рівномірно розподіленої на [0,1] випадкової величини та підраховується значення $\pi_1 = nx_1/T_N$, де $T_N = \sum_{k\in U} x_k$. Якщо $\varepsilon_1 < \pi_1$, то елемент 1 потрапляє у вибірку, та не потрапляє – в іншому випадку.

3) Для кожного наступного елемента $k = 2, 3, ...$ незалежно генерується значення ε_k рівномірно розподіленої на [0,1] випадкової величини та підраховується значення $\pi'_k = \frac{(n-n_k)x_k}{t_k}$, де n_k – кількість елементів, що потрапили у вибірку, серед перших $k-1$ елементів, $t_k = x_k + x_{k+1} + ... + x_N$.

Якщо $\varepsilon_k < \pi_k'$, то елемент k потрапляє у вибірку, та не потрапляє – в іншому випадку.

4) Процедура відбору з пунктів 2) та 3) закінчується, коли $n_k = n$ або $k = k^*$, де $k^* = \min\{k_0, N - n + 1\}$, k_0 – найменше з тих k, для яких $nx_k/t_k \geq 1$.

5) Якщо $n_{k^*} < n$, то в результаті застосування процедури відбору з пунктів 2) та 3) ми не отримали необхідної кількості елементів у вибірку. Ті $n - n_{k^*}$ елементів, що залишилось вибрати з $N - k^* + 1$ потрапляють у вибірку за допомогою схеми відбору 1, що наведена в підрозділі 2.1. А саме: для елемента $k \geq k^*$ генерується значення ε_k рівномірно розподіленої на [0,1] випадкової величини та порівнюється зі значенням $\pi_k'' = \frac{n-n_k}{N-k+1}$. Якщо $\varepsilon_k < \pi_k''$, то елемент k потрапляє у вибірку, та не потрапляє – в іншому випадку.

Процедура завершується, коли $n_k = n$.

У результаті застосування цієї схеми відбору ймовірності включення будуть такими:

$$\pi_k = \begin{cases} \frac{nx_k}{T_N}, & k = 1, ..., k^* - 1; \\ \frac{nT_{k^*}}{T_N(N-k^*+1)}, & k = k^*, ..., N. \end{cases}$$

Для підрахунку ймовірностей включення другого порядку введемо величини: $g_1 := \frac{1}{t_2}$, $g_k = g_{k-1}\frac{t_k - x_{k-1}}{t_{k+1}}$, $k = 2, 3, ..., k^* - 1$. Тоді

$$\pi_{kl} = \begin{cases} \frac{n(n-1)}{T_N} x_k x_l g_k, & 1 \leq k < l < k^*; \\ \frac{n(n-1)}{T_N} \frac{t_{k^*}}{(N-k^*+1)} x_k g_k, & 1 \leq k < k^* \leq l \leq N; \\ \frac{n(n-1)}{T_N} \frac{\left(\frac{t_{k^*}}{(N-k^*+1)}\right)^2 (t_{k^*} - x_{k^*-1})}{t_{k^*} - \frac{t_{k^*}}{(N-k^*+1)}} g_{k^*-1}, & k^* \leq k < l \leq N. \end{cases}$$

6.3. Вправи та питання для самоконтролю

6.1. Чим відрізняється відбір Пуассона від відбору Бернуллі?

6.2. Чи є нерівноймовірнісний відбір більш ефективним, ніж відбір з рівними ймовірностями? В якому випадку?

6.3. Яку оцінюючу статистику використовують при відборі, p-пропорційному до розміру, та при відборі, π-пропорційному до розміру?

6.4. За допомогою відбору Пуассона із середнім розміром вибірки 10 отримана вибірка з генеральної сукупності, що складається зі 100 елементів, для оцінювання сумарного T характеристики y. Ймовірності, з якими елементи потрапляли у вибірку, вибирались пропорційно до змінної x. Таким чином було обрано 12 елементів. Результати обстеження наведено в таблиці:

k	1	2	3	4	5	6	7	8	9	10	11	12
x_k	54	671	2	27	29	62	4	48	33	446	12	46
y_k	5,2	59,8	2,2	2,5	2,9	6,8	3,7	4,2	4,1	38,9	1,1	4,8

Знайти:

1) оцінку Горвіца–Томпсона для сумарного T характеристики y, якщо $\sum_{k \in U} x_k = 8\,182$;

2) незміщену оцінку для дисперсії оцінки Горвіца–Томпсона;

3) оцінку коефіцієнта варіації для $\widehat{t}_\pi$.

6.5. Для оцінювання середнього $\overline{Y}$ характеристики y генеральної сукупності розміру 4 застосовано відбір, π-пропорційний до розміру. Для цього було вибрано два елементи пропорційно до змінної x за схемою Брюера: $y_1 = 65$, $y_4 = 22$. Відомо, що $x_1 = 67$, $x_2 = 14$, $x_3 = 45$, $x_4 = 24$. Обчислити:

1) незміщену оцінку для середнього;

2) незміщену оцінку для дисперсії цієї оцінки;

3) оцінку для коефіцієнта варіації для $\widehat{\overline{y}}_\pi$: $cv(\widehat{\overline{y}}_\pi) = \frac{\sqrt{\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi)}}{\widehat{\overline{y}}_\pi}$.

6.6. Нехай для кожного елемента генеральної сукупності, що складається з десяти елементів, відомі значення допоміжної змінної x:

$$10, 2, 8, 6, 2, 10, 1, 1, 6, 4.$$

Утворити вибірку з $n = 4$ елементів, використовуючи схему Сантера. Для знаходжень значень ε_k можна скористатись таблицею випадкових чисел, наведеною в додатку 2. Для цього числа з цієї таблиці можна розглядати, як знаки після коми у рівномірно розподіленої на [0,1] випадкової величини.

6.7. Для генеральної сукупності з шести елементів відомі значення допоміжної характеристики x:

$$x_1 = 400,\ x_2 = x_3 = 15,\ x_4 = 10,\ x_5 = x_6 = 5.$$

1) Отримати вибірку, що складається з трьох елементів з імовірностями, π-пропорційними до x, використовуючи метод Сантера.

   При відборі скористатися такими значеннями для рівномірно розподіленої на [0,1] випадкової величини:

   $$\varepsilon_1 = 0,28; \varepsilon_2 = 0,37; \varepsilon_3 = 0,95;\ \varepsilon_4 = 0,48; \varepsilon_5 = 0,83; \varepsilon_6 = 0,74.$$

2) Підрахувати ймовірності включення другого порядку π_{23} та π_{24}.

6.8. Метод Мідзуно [11]. Нехай розмір генеральної сукупності $N \geq 3$. Розглянемо такий метод відбору з нерівними ймовірностями: перший елемент потрапляє у вибірку з нерівними ймовірностями p_k ($\sum_{k \in U} p_k = 1$). Інші $n - 1$ елементи потрапляють у вибірку з тих $N - 1$, що залишились, за допомогою ПВВбП.

1) Знайти ймовірності включення першого та другого порядку.

2) Виразити ймовірності включення другого порядку через ймовірності включення першого порядку.

3) Чи задовольняють ці ймовірності умову Єйтса–Гранді–Сена?

Розділ 7

Стратифікований відбір

7.1. Означення та застосування стратифікованого відбору

При *стратифікованому відборі (СТВ)* генеральна сукупність ділиться на підсукупності, що не перетинаються. Ці підсукупності називаються *стратами* (англ. `stratum`) [4]. Процедура поділу генеральної сукупності на страти називається *стратифікацією.* Вибіркові одиниці всередині страт однорідні (подібні) за певною ознакою чи ознаками.

У кожній страті проводиться відбір елементів. Метод відбору елементів зі страти може бути будь-яким, причому він необов'язково однаковий для всіх страт. Відбір з однієї страти не залежить від відбору з іншої страти.

Генеральну сукупність можна стратифікувати за будь-якою характеристикою (змінною), наявною у кожної одиниці з вибіркової основи до проведення відбору (наприклад, вік, стать, район проживання, дохід тощо). Змінна, за якою проводиться стратифікація, називається *стратифікаційною.* Стратифікаційних змінних може бути відразу кілька, але зазвичай для стратифікації вибирають одну чи дві змінні.

Чому виникає потреба використовувати стратифікацію? Для цього є багато причин, головною з яких є те, що завдяки стратифікації стратегія відбору стає більш ефективною. Чим більше змінюється досліджувана характеристика від однієї вибіркової одиниці до іншої, тим більший розмір вибірки потрібен для отримання достатньо точної оцінки. Але якщо створити страти, всередині яких значення досліджуваної характеристики (наприклад,

[4] «Stratum» перекладається з англійської як «шар», тож іноді цей метод відбору називають «відбором з розшаруванням». Але в нечисленній україномовній літературі з вибіркових обстежень частіше вживаються назви «страта» та «стратифікований відбір», тож саме ці назви і будемо використовувати в цьому посібнику.

доходу) схожі, а для різних страт вони суттєво відрізняються, то малої вибірки з кожної страти буде цілком достатньо для знаходження точної оцінки для страти. Комбінуючи оцінки, отримані для страт, матимемо оцінку досліджуваної характеристики для всієї генеральної сукупності, при цьому сумарний розмір вибірки при стратифікованому відборі буде значно меншим, ніж розмір простої випадкової вибірки, потрібний для такої самої точності оцінювання.

Наведемо ще кілька причин використання стратифікованого відбору:

1) Припустимо, що необхідно отримати оцінки досліджуваної характеристики для певних *областей дослідження* – підсукупностей генеральної сукупності, і задано бажану точність цих оцінок. Якщо за вибірковою основою можна визначити належність одиниці генеральної сукупності до тієї чи іншої області дослідження, то кожну таку область можна розглядати як окрему страту. Далі з кожної страти можна відібрати ймовірнісну вибірку такого обсягу, який буде гарантувати задану точність оцінювання.

2) При проведенні вибіркового обстеження практичні аспекти, які стосуються невідповідей, складності вимірювань та обсягу додаткової інформації, можуть значно відрізнятися для різних підсукупностей генеральної сукупності. Наприклад, не всі категорії населення однаково охоче відповідають на запитання, що стосуються їхніх прибутків. Це означає, що в деяких ситуаціях для того, щоб підвищити ефективність обстеження, варто для кожної страти окремо підбирати методи відбору та оцінювання згідно з її особливостями.

3) Мета вибіркового обстеження може вимагати поділу певної географічної території на окремі райони (наприклад, потрібно оцінити рівень безробіття не тільки для України в цілому, а й для кожної області окремо). Або з точки зору практичного втілення обстеження генеральна сукупність адміністративно розбита на регіони, в кожному з яких обстеження проводиться своїм підрозділом. У таких випадках цілком природно розглядати кожен район як окрему страту.

Зауваження 7.1. При стратифікованому відборі страта, по суті, стає окремою популяцією, а це означає, що для кожної страти потрібно знайти необхідний розмір вибірки.

Зауваження 7.2. Найбільш корисною стратифікація є тоді, коли стратифікаційна змінна, по-перше, тісно пов'язана з метою дослідження (тобто, з досліджуваною характеристикою) і, по-друге, її можна легко обстежити або виміряти.

Для того, щоб максимально використати переваги стратифікованого відбору і отримати стратифіковану вибірку, яка буде оптимальною як з точки зору ефективності, так і з точки зору практичної реалізації, потрібно визначитися щодо кількох технічних питань.

1) Поділ генеральної сукупності на страти:

   - якщо є можливість вибору, то за якою змінною стратифікувати (наприклад, за віком, статтю чи фахом)?
   - скільки страт має бути (наприклад, у випадку стратифікації за віком – скільки має бути вікових груп)?
   - як саме розмежовувати страти (наприклад, якщо вік – стратифікаційна змінна, то якими повинні бути межі вікових груп)?

2) Методи відбору та оцінювання:

   - для кожної страти потрібно підібрати вибірковий дизайн та визначити необхідний розмір вибірки; часто один і той самий метод відбору використовується для всіх страт;
   - для кожної страти потрібно вибрати метод оцінювання; знову ж таки, часто однаковий метод оцінювання використовується для всіх страт.

Всі ці питання взаємопов'язані. Остаточне рішення залежить від багатьох обставин, у тому числі й від вимог щодо точності оцінок для різних змінних і областей дослідження, вартості обстеження, адміністративних обмежень тощо.

7.2. π-оцінка сумарного значення при стратифікованому відборі

Нехай генеральна сукупність $U = \{1, \dots, k, \dots, N\}$ поділена на H страт: $U_1, \dots, U_h, \dots, U_H$, де $U_h = \{k \in U : k$ належить h-й страті$\}$. При стратифікованому відборі ймовірнісна вибірка s_h вибирається зі страти U_h згідно з вибірковим дизайном $p_h(\cdot)$, $h = 1, \dots, H$, при цьому відбір з будь-якої страти проводиться незалежно від відборів з решти страт.

Результуюча вибірка s є об'єднанням вибірок, відібраних зі страт, тобто

$$s = s_1 \cup s_2 \cup \dots \cup s_H.$$

Внаслідок незалежності відборів зі страт матимемо

$$p(s) = p_1(s_1)p_2(s_2)\dots p_H(s_H).$$

Будемо вважати, що кількість елементів в h-й страті відома. Позначимо її через N_h. Оскільки страти утворюють розбиття генеральної сукупності, то

$$N = \sum_{h=1}^{H} N_h.$$

Очевидно, що сумарне значення досліджуваної характеристики для всієї генеральної сукупності є сумою сумарних значень у стратах:

$$T = \sum_{k \in U} y_k = \sum_{h=1}^{H} T_h = \sum_{h=1}^{H} N_h \overline{Y}_h, \tag{7.1}$$

де $T_h = \sum\limits_{k \in U_h} y_k$ – сумарне значення в h-й страті, а $\overline{Y}_h$ – середнє значення в h-й страті.

Позначимо через $W_h = \frac{N_h}{N}$ – вагу страти U_h. Тоді середнє для генеральної сукупності матиме вигляд

$$\overline{Y} = \sum_{h=1}^{H} W_h \overline{Y}_h. \tag{7.2}$$

***Твердження* 7.1.** *При стратифікованому відборі π-оцінка сумарного значення $T = \sum\limits_{k \in U} y_k$ має вигляд*

$$\widehat{t}_\pi = \sum_{h=1}^{H} \widehat{t}_{h\pi}, \tag{7.3}$$

де $\widehat{t}_{h\pi}$ – це π-оцінка сумарного значення в h-й страті, $T_h = \sum\limits_{k \in U_h} y_k$. Дисперсія оцінки $\widehat{t}_\pi$ при стратифікованому відборі така:

$$\mathcal{D}_{\text{СТВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} \mathcal{D}_h(\widehat{t}_{h\pi}), \tag{7.4}$$

де $\mathcal{D}_h(\widehat{t}_{h\pi})$ – дисперсія оцінки $\widehat{t}_{h\pi}$. Незміщена оцінка дисперсії оцінки $\widehat{t}_\pi$ обчислюється за формулою

$$\widehat{\mathcal{D}}_{\text{СТВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} \widehat{\mathcal{D}}_h(\widehat{t}_{h\pi}), \tag{7.5}$$

за умови, що незміщена оцінка $\widehat{\mathcal{D}}_h(\widehat{t}_{h\pi})$ дисперсії $\mathcal{D}_h(\widehat{t}_{h\pi})$ існує для всіх $h = \overline{1, H}$.

Доведення. З означення стратифікованого відбору випливає, що

$$\pi_k = P(k \in s) = P(k \in s_h).$$

Таким чином, π-оцінка сумарного значення T має вигляд

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum_{k \in s_h} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \widehat{t}_{h\pi},$$

де $\widehat{t}_{h\pi}$ – π-оцінка сумарного значення в h-й страті. Оскільки відбори зі страт проводяться незалежно, то випадкові величини $\widehat{t}_{h\pi}$, $h = 1, \ldots, H$, – незалежні. Звідси згідно з властивостями дисперсії отримуємо формули (7.4) та (7.5). □

Зазвичай один і той самий метод відбору використовується для всіх страт. Якщо в кожній страті для побудови вибірки використовується простий випадковий відбір без повернення, то такий метод відбору одиниць з генеральної сукупності називається *стратифікованим простим випадковим відбором (СТПВВ)*. Якщо в кожній страті для побудови вибірки використовується систематичний відбір, то такий метод відбору називається *стратифікованим систематичним відбором (СТСВ)*. Аналогічно, *стратифікований відбір Бернуллі (СТВБ)* означає, що для побудови стратифікованої вибірки в кожній страті використовується відбір Бернуллі, і т. д. Використовуючи відомості попередніх розділів, із твердження 7.1 можна легко вивести формули для оцінювання сумарного значення досліджуваної характеристики, дисперсії та оцінки дисперсії отриманої оцінки для кожного з цих випадків.

Зокрема, для СТПВВ матимемо таке твердження.

Твердження 7.2. *Нехай n_h – фіксований розмір простої випадкової вибірки з h-ї страти, $h = 1, \ldots, H$. При СТПВВ π-оцінка сумарного значення $T = \sum\limits_{k \in U} y_k$ має вигляд*

$$\widehat{t}_\pi = \sum_{h=1}^{H} N_h \overline{y}_h, \tag{7.6}$$

де $\overline{y}_h = \sum\limits_{k \in s_h} \frac{y_k}{n_h}$ – вибіркове середнє в h-й страті.

Дисперсія оцінки $\widehat{t}_\pi$ при СТПВВ запишеться так:

$$\mathcal{D}_{\text{СТПВВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_h^2, \tag{7.7}$$

де $f_h = n_h / N_h$ – частка відбору з h-ї страти,
$S_h^2 = \frac{1}{N_h - 1} \sum\limits_{k \in U_h} (y_k - \overline{Y}_h)^2$ – дисперсія в h-й страті,
$\overline{Y}_h = \sum\limits_{k \in U_h} \frac{y_k}{N_h}$ – середнє в h-й страті, $h = 1, \ldots, H$.

Незміщена оцінка дисперсії оцінки $\widehat{t}_\pi$ обчислюється за формулою

$$\widehat{\mathcal{D}}_{\text{СТПВВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} \widehat{S}_h^2, \tag{7.8}$$

де $\widehat{S}_h^2 = \frac{1}{n_h - 1} \sum\limits_{k \in s_h} (y_k - \overline{y}_h)^2$ – вибіркова дисперсія в h-й страті.

7.3. Оптимальне розміщення стратифікованої вибірки

Розглянемо стратифіковану генеральну сукупність і припустимо, що для кожної страти вибрано відповідний метод відбору. Для оцінювання сумарного значення досліджуваної характеристики популяції планується використати π-оцінку. Перш ніж розпочинати безпосередній відбір елементів, потрібно ще визначити необхідні розміри вибірок зі страт: $n_h, h = 1, \ldots, H$.

Нехай стратифікований вибірковий дизайн такий, що дисперсію π-оцінки сумарного значення генеральної сукупності можна подати у вигляді:

$$\mathcal{D}_{\text{СТВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} \frac{A_h}{n_h} + B =: D, \tag{7.9}$$

де вирази A_h та B не залежать від n_h. Зокрема, в такому вигляді можна записати дисперсію π-оцінки сумарного значення популяції при СТПВВ та СТВБ.

Припустимо, що загальні витрати на вибіркове обстеження можна зобразити у вигляді лінійної функції

$$C = c_0 + \sum_{h=1}^{H} n_h c_h, \tag{7.10}$$

де c_0 – деякі фіксовані витрати, а $c_h > 0$ – витрати на обстеження одного елемента h-ї страти.

Задачу *оптимального розміщення вибірки*[5] можна сформулювати так: необхідно знайти такі значення розмірів вибірок зі страт $n_h, h = 1, \ldots, H$, які будуть мінімізувати дисперсію D при фіксованих витратах C, або, навпаки, які будуть мінімізувати витрати C при фіксованому значенні дисперсії D.

***Твердження* 7.3.** *Якщо стратифікований вибірковий дизайн такий, що дисперсію $\mathcal{D}_{\text{СТВ}}(\widehat{t_\pi})$ можна записати у вигляді (7.9), то оптимальне розміщення при лінійній функції витрат виду (7.10) досягається тоді, коли n_h пропорційне $(A_h/c_h)^{1/2}$.*

Доведення. Позначимо $D^* = D - B$ та $C^* = C - c_0$.

У цьому випадку задача знаходження оптимального розміщення еквівалентна задачі мінімізації добутку

$$D^*C^* = \left(\sum_{h=1}^{H} \frac{A_h}{n_h}\right)\left(\sum_{h=1}^{H} n_h c_h\right). \tag{7.11}$$

Застосуємо нерівність Коші:

$$\left(\sum a_h^2\right)\left(\sum b_h^2\right) \geq \left(\sum a_h b_h\right)^2,$$

в якій покладемо $a_h = (A_h/n_h)^{1/2}$ та $b_h = (n_h c_h)^{1/2}$. Тоді

$$D^*C^* \geq \left[\sum_{h=1}^{H} (A_h c_h)^{1/2}\right]^2.$$

Рівність у цій нерівності (тобто, мінімальне значення добутку D^*C^*) досягається тоді і тільки тоді, коли b_h/a_h – стала величина для всіх $h = 1, \ldots, H$, тобто, при

$$\left(\frac{n_h c_h}{A_h/n_h}\right)^{1/2} = \text{const}$$

[5] Іноді замість «оптимального розміщення» вживають термін «оптимальний розподіл» вибірки.

або

$$n_h \propto (A_h/c_h)^{1/2}.$$

□

***Твердження* 7.4.** *1) Мінімізуючи дисперсію D виду (7.9) при фіксованих витратах C виду (7.10), отримаємо оптимальні розміри вибірок зі страт:*

$$n_h = \frac{(C - c_0)(A_h/c_h)^{1/2}}{\sum\limits_{h=1}^{H} (A_h c_h)^{1/2}}, \quad h = 1, \ldots, H. \tag{7.12}$$

Мінімальне значення дисперсії при цьому становить

$$\mathcal{D}_{\text{опт}} = \frac{1}{(C - c_0)}\left[\sum_{h=1}^{H} (A_h c_h)^{1/2}\right]^2 + B. \tag{7.13}$$

2) Мінімізуючи витрати C при фіксованій дисперсії D, отримаємо таке оптимальне розміщення:

$$n_h = \frac{(A_h/c_h)^{1/2}\left[\sum\limits_{h=1}^{H} (A_h c_h)^{1/2}\right]}{(D - B)}, \quad h = 1, \ldots, H. \tag{7.14}$$

При цьому мінімальне значення функції витрат дорівнює

$$C_{\text{опт}} = c_0 + \frac{1}{D - B}\left[\sum_{h=1}^{H} (A_h c_h)^{1/2}\right]^2. \tag{7.15}$$

Зауваження 7.3. У твердженнях 7.3 та 7.4 неявно припускається, що розв'язок задачі знаходження оптимального розміщення задовольняє умову $n_h \leq N_h$ для кожного $h = 1, \ldots, H$. Якщо ж одне або кілька з цих обмежень не виконуються, то для відповідних страт потрібно покласти $n_h = N_h$, а для решти страт перерахувати (підправити) розміри вибірок. Наприклад, розглянемо випадок, коли дисперсія D фіксована, а витрати C потрібно мінімізувати. Нехай n_h перевищують N_h при $h = 1, \ldots, p$, $p \leq H$ (ми

завжди можемо змінити нумерацію страт потрібним чином). Тоді нові розміри вибірок зі страт обчислюємо так:

$$\begin{cases} n_h = N_h, & h = 1, \dots, p; \\ n_h = K(A_h/c_h)^{1/2}, & h = p+1, \dots, H, \end{cases}$$

де стала K визначається за формулою

$$K = \sum_{h=p+1}^{H} (A_h c_h)^{1/2} \Big/ \left(D - B - \sum_{h=1}^{p} A_h/N_h \right).$$

Якщо після перерахунку всі $n_h \le N_h$, то це і буде шуканий розв'язок. Якщо ж ні, то наведену процедуру повторюємо доти, поки ця умова не буде виконана.

Зауваження 7.4. При визначенні розмірів вибірок зі страт слід пам'ятати і про такі дві очевидні умови:

1) для обчислення вибіркових середніх у стратах потрібно, щоб $n_h \ge 1$, $h = 1, \dots, H$;

2) для обчислення вибіркових дисперсій потрібно, щоб $n_h \ge 2$, $h = 1, \dots, H$.

Зауваження 7.5. Існує багато ефективних критеріїв стратифікації, які можна використовувати при вибірковому обстеженні. Іноді дослідники працюють з дуже великою кількістю страт. Трапляються випадки, коли відбирають тільки по одному елементу з кожної страти. Звичайно, в таких випадках не можна використовувати оцінку дисперсії (7.5). Натомість пропонується так званий *метод вироджених страт*, який полягає в такому: страти розбивають на пари, а потім на основі двох спостережень у кожній парі оцінюється дисперсія. Нехай оцінка сумарного значення генеральної сукупності $T = \sum_{h=1}^{H} T_h$ дорівнює $\widehat{t}_\pi = \sum_{h=1}^{H} \widehat{t}_h$, де $\widehat{t}_h = y_k/\pi_k$ – незміщена оцінка сумарного значення T_h в h-й страті, $h = 1, \dots, H$, а елемент k відбирається з відповідної страти з імовірністю π_k. Наприклад, при простому випадковому відборі з

h-ї страти $\pi_k = 1/N_h$ для всіх елементів цієї страти. Припустимо, що H – парне число. Позначимо індексами $j1$ та $j2$ страти у j-й парі, $j = 1, \dots, M = H/2$. Тоді оцінка дисперсії при застосуванні методу вироджених страт матиме вигляд

$$\widehat{\mathcal{D}}_{\text{вир}} = \sum_{j=1}^{M} (\widehat{t}_{j1} - \widehat{t}_{j2})^2.$$

Математичне сподівання цієї оцінки дорівнює

$$E(\mathcal{D}_{\text{вир}}) = \mathcal{D}(\widehat{t}_\pi) + \sum_{j=1}^{M} (t_{j1} - t_{j2})^2,$$

тобто $\widehat{\mathcal{D}}_{\text{вир}}$ зазвичай переоцінює дисперсію $\mathcal{D}(\widehat{t}_\pi)$ і призводить до консервативних довірчих інтервалів. Якщо страти можна розбити на пари так, щоб $\forall j = 1, \dots, M = H/2$ мала місце рівність $\widehat{t}_{j1} = \widehat{t}_{j2}$, то оцінка $\widehat{\mathcal{D}}_{\text{вир}}$ буде незміщеною. У більшості випадків важко уникнути зміщення цієї оцінки, оскільки для цього потрібно знати сумарні значення в стратах.

Приклад 7.1. *Оптимальне розміщення при СТПВВ.* У випадку стратифікованого простого випадкового відбору дисперсію $\mathcal{D}_{\text{СТПВВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_h^2$ можна записати у вигляді формули (7.9) з $A_h = N_h^2 S_h^2$, $B = -\sum_{h=1}^{H} N_h S_H^2$. Мінімізуючи дисперсію при фіксованих витратах, отримаємо

$$n_h = (C - c_0) \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h \sqrt{c_h}}. \tag{7.16}$$

Із виразу (7.16) випливає, що чим більша дисперсія в страті, тим більше n_h. Аналогічно, чим менші витрати на обстеження елемента зі страти, тим більший розмір вибірки зі страти. ◊

7.4. Альтернативні розміщення при СТПВВ

У цьому підрозділі більш детально розглянемо стратифікований простий випадковий відбір та альтернативні розміщення стратифікованої вибірки, які можна використовувати в даному випадку.

Припустимо, що витрати на обстеження одного елемента однакові для всіх страт, тобто $c_1 = c_2 = \ldots = c_H$.

Позначимо через n сумарний розмір вибірки: $n = \sum_{h=1}^{H} n_h$.

7.4.1. Розміщення Неймана

При стратифікованому простому випадковому відборі оптимальне розміщення (7.16), яке мінімізує дисперсію при фіксованих загальних витратах за умови, що витрати на обстеження одного елемента однакові для всіх страт, набуде вигляду:

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}, \quad h = 1, \ldots, H. \tag{7.17}$$

Таке розміщення також називають *розміщенням Неймана*. Для знаходження оптимальних значень n_h за формулою (7.17) потрібно знати стандартні відхилення в стратах, тобто S_h, $h = 1, \ldots, H$. Ми тільки тоді зможемо отримати оптимальне розміщення (тобто, мінімальне значення дисперсії), коли матимемо точні значення S_h. На практиці це неможливо. Але при повторному обстеженні можна використати результати попереднього обстеження для того, щоб отримати досить близьке наближення до істинних значень стандартних відхилень у стратах. Тоді отримаємо розміщення, близьке до оптимального.

7.4.2. Пропорційне розміщення

Пропорційне розміщення визначається за формулою

$$n_h = \frac{nN_h}{N}, \quad h = 1, \ldots, H. \tag{7.18}$$

Оскільки ми вважаємо, що розміри страт відомі заздалегідь, то таке розміщення завжди можна обчислити.

Якщо у всіх стратах однакові стандартні відхилення ($S_1 = S_2 = \ldots = S_H$), то оптимальне розміщення Неймана (7.17) набуде виду (7.18), тобто в цьому випадку пропорційне розміщення буде оптимальним. В інших випадках пропорційне розміщення дає більші значення дисперсії, ніж розміщення Неймана, особливо тоді, коли значення S_h сильно відрізняються.

7.4.3. Розміщення, пропорційне до сумарного значення змінної y

Розміщення, пропорційне до сумарного значення змінної y, визначається так:

$$n_h = n\frac{\sum_{k\in U_h} y_k}{\sum_{k\in U} y_k} = n\frac{T_{yh}}{T_y}, \quad h = 1, \ldots, H, \tag{7.19}$$

при цьому припускається, що змінна y набуває тільки додатних значень.

Це розміщення є оптимальним, тобто співпадає з розміщенням Неймана (7.17) тоді, коли коефіцієнти варіації $CV_h = \frac{S_h}{\overline{Y}_h}$ однакові для всіх страт.

Оскільки сумарні значення змінної y як для страт, так і для всієї генеральної сукупності заздалегідь невідомі, то безпосереднє застосування розміщення (7.19) на практиці неможливе.

7.4.4. x-оптимальне розміщення

Розглянемо x-оптимальне розміщення, яке є альтернативою розміщенню (7.17) і має переваги з точки зору практичного застосування, тому часто використовується на практиці.

Нехай x – додаткова змінна, корельована зі змінною y. Позначимо через S_{xh} стандартне відхилення змінної x в h-й страті, $h = \overline{1, H}$. Значення S_{xh} відомі заздалегідь для всіх $h = \overline{1, H}$.

Тоді *x-оптимальне розміщення* визначається так:

$$n_h = n\frac{N_h S_{xh}}{\sum\limits_{h=1}^{H} N_h S_{xh}}, \quad h = 1, \ldots, H. \tag{7.20}$$

Якщо змінні x та y «ідеально» корельовані ($y_k = a + bx_k$, $k = 1, \ldots, N$), то розміщення (7.20) фактично є оптимальним. Якщо кореляція сильна, але не ідеальна, то розміщення (7.20) є близьким до оптимального.

7.4.5. Розміщення, пропорційне до сумарного значення змінної x

Розміщення, пропорційне до сумарного значення змінної x, визначається за виразом

$$n_h = n\frac{\sum\limits_{k\in U_h} x_k}{\sum\limits_{k\in U} x_k} = n\frac{T_{xh}}{T_x}, \quad h = 1, \ldots, H, \tag{7.21}$$

при цьому припускається, що додаткова змінна x завжди набуває додатних значень, а сумарні значення в стратах $\sum\limits_{k\in U_h} x_k$ відомі заздалегідь. Таке розміщення легко обчислюється та широко застосовується на практиці. Якщо змінні x та y сильно корельовані, а коефіцієнти варіації приблизно однакові для всіх страт, то розміщення (7.21) буде близьким до оптимального.

7.5. Порівняння дисперсій π-оцінки сумарного значення при оптимальному та пропорційному розміщеннях

При стратифікованому простому випадковому відборі

$$\widehat{t}_\pi = \sum_{h=1}^{H} N_h \overline{y}_h = N \sum_{h=1}^{H} W_h \overline{y}_h.$$

При оптимальному розміщенні Неймана з формули (7.17) отримаємо

$$\mathcal{D}^{\text{опт}}_{\text{СТПВВ}}(\widehat{t}_\pi) = \frac{N^2}{n}\left(\sum_{h=1}^{H} W_h S_h\right)^2 - N\sum_{h=1}^{H} W_h S_h^2. \tag{7.22}$$

У випадку пропорційного розміщення (7.18) будемо мати

$$\mathcal{D}^{\text{проп}}_{\text{СТПВВ}}(\widehat{t}_\pi) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{H} W_h S_h^2. \tag{7.23}$$

Оскільки розміщення Неймана – оптимальне, то дисперсія (7.22) не може перевищувати дисперсію (7.23). Рівність має місце тільки тоді, коли дисперсії в стратах однакові. Зазвичай дисперсія (7.22) значно менша, ніж (7.23), особливо у випадку, коли дисперсії в стратах суттєво відрізняються.

7.6. Порівняння дисперсій π-оцінки сумарного значення при СТПВВ та ПВВбП

Якщо рішення щодо розміщення стратифікованої вибірки математично обґрунтоване, то при однаковому розмірі вибірки n стратифікований простий випадковий відбір зазвичай є більше ефективним методом відбору, ніж простий випадковий. Для прикладу порівняємо дисперсії π-оцінки сумарного значення при простому випадковому відборі без повернення та при стратифікованому простому випадковому відборі з пропорційним розміщенням.

При простому випадковому відборі $\widehat{t}_\pi = N\overline{y} = N\sum\limits_{k\in s} y_k/n$. Використовуючи формули для обчислення дисперсій цієї оцінки при ПВВбП та при СТПВВ з пропорційним розміщенням, можна легко показати, що

$$\mathcal{D}_{\text{ПВВбП}}(N\overline{y}) - \mathcal{D}^{\text{проп}}_{\text{СТПВВ}}\left(\sum_{h=1}^{H} N_h \overline{y}_h\right) = \tag{7.24}$$

$$= \frac{N^3}{N-1}\left(\frac{1}{n} - \frac{1}{N}\right)\left[\sum_{h=1}^{H} W_h(\overline{Y}_h - \overline{Y})^2 - \frac{1}{N}\sum_{h=1}^{H}(1 - W_h)S_h^2\right],$$

де $\overline{Y}_h = \sum\limits_{k\in U_h} y_k/N_h$ – середнє в h-й страті, $h = 1,\dots,H$, а $\overline{Y} = \sum\limits_{k\in U} y_k/N$ – середнє для всієї генеральної сукупності.

Із виразу (7.24) випливає, що теоретично дисперсія при СТПВВ з пропорційним розміщенням може трохи перевищувати дисперсію при ПВВбП, а саме тоді, коли середні в стратах однакові або майже однакові. Причиною цього може бути невдалий поділ на страти. Якщо ж стратифікація дійсно потрібна і правильно виконана, то доданок $\sum\limits_{h=1}^{H} W_h(\overline{Y}_h - \overline{Y})^2$ буде значно більшим, ніж $N^{-1}\sum\limits_{h=1}^{H}(1-W_h)S_h^2$, і стратифікований простий випадковий відбір з пропорційним розміщенням буде значно ефективнішим порівняно з простим випадковим відбором без повернення.

7.7. Вправи та питання для самоконтролю

7.1. Яка вибірка краще відображає структуру генеральної сукупності: стратифікована чи проста випадкова? Чому?

7.2. У чому полягає задача оптимізації розміщення стратифікованої вибірки?

7.3. Чим відрізняється розміщення Неймана від оптимального розміщення стратифікованої вибірки?

7.4. Який відбір і в яких ситуаціях є більш ефективним: СТПВВ чи ПВВбП?

7.5. Визначити ймовірності включення π_k та π_{kl} у випадку стратифікованого простого випадкового відбору.

7.6. Довести твердження 7.2.

7.7. Вивести формули для обчислення оцінки сумарного значення, дисперсії цієї оцінки, а також для оцінювання дисперсії отриманої оцінки при стратифікованому систематичному відборі та стратифікованому відборі Бернуллі.

7.8. У деякому вибірковому обстеженні використовується СТПВВ. Припустимо, що потрібно оцінити середнє значення $\overline{Y}$ досліджуваної характеристики генеральної сукупності з такою точністю, щоб умова

$$\left|\sum_{h=1}^{H} W_h\overline{y}_h - \overline{Y}\right| \le a$$

виконувалася при заданому значенні сталої a щонайменше з імовірністю $1-\alpha$. Довести, що розмір вибірки, необхідний для виконання цієї умови, задовольняє нерівність

$$n \ge \frac{\left(\frac{z}{a}\right)^2 \sum\limits_{h=1}^{H} \frac{W_h^2 S_h^2}{w_h}}{1 + \frac{1}{N}\left(\frac{z}{a}\right)^2 \sum\limits_{h=1}^{H} W_h^2 S_h^2},$$

де $w_h = n_h/n$, $z = z_{1-\alpha/2}$.

7.9. Нехай у попередній вправі $N = 1000$, $H = 2$, $W_1 = 1 - W_2 = 0,8$, $S_1^2 = 4$, $S_2^2 = 16$, $z = 1,96$, $a = 0,5$. Дослідити, як змінюється необхідний розмір вибірки n як функція від $w_1 = 1 - w_2$, якщо w_1 змінюється від 0 до 1.

7.10. Довести твердження 7.4.

Розділ 8

Кластерний, двостадійний та багатостадійний відбір

8.1. Основні поняття

У попередніх розділах ми розглядали вибіркові дизайни, в яких елементи відбиралися з генеральної сукупності відразу, тобто за одну стадію відбору. Але безпосередній відбір елементів до вибірки використовують не завжди. Однією з причин цього може бути відсутність вибіркової основи з повним переліком елементів генеральної сукупності, а створення такої основи є надто коштовним. Іншою причиною може бути територіальне розташування елементів генеральної сукупності: вибіркові одиниці можуть бути розташовані далеко одна від одної, а це призводить до збільшення як фінансових витрат (переїзд від однієї одиниці до іншої), так і тривалості обстеження. Також у цьому випадку важко контролювати якість обстеження, внаслідок чого можна отримати високий рівень невідповідей або значні похибки вимірювань.

Кластерний відбір (англ. `cluster sampling`) є одним з методів відбору, які можна використовувати тоді, коли безпосередній відбір елементів з популяції неможливий або небажаний. Для цього елементи генеральної сукупності об'єднують у групи, які називаються *кластерами*. Для кластерного відбору будемо використовувати скорочення КВ.

У випадку *одностадійного кластерного відбору* (ОКВ) спочатку відбирають кластери, а потім обстежують всі елементи відібраних кластерів. Наприклад, при обстеженні населення міста кластером може бути сім'я, домогосподарство або мешканці одного житлового будинку. Спочатку будують вибірку кластерів, наприклад, домогосподарств, а потім опитують всіх членів тих домогосподарств, що потрапили до вибірки.

У випадку *двостадійного відбору* (ДВ) популяцію ділять на кластери – первинні вибіркові одиниці (ПВО), які складаються з

окремих елементів або з дрібніших груп (кластерів) елементів – вторинних вибіркових одиниць (ВВО). На першій стадії відбору отримують імовірнісну вибірку первинних вибіркових одиниць. На другій стадії відбору з тих первинних вибіркових одиниць, що потрапили до вибірки на першій стадії, відбирають вторинні вибіркові одиниці (елементи або кластери елементів). Після цього обстежують відібрані ВВО. Якщо ВВО – це кластери, то обстежують всі елементи відібраних ВВО. Якщо всі вторинні вибіркові одиниці – це окремі елементи, то такий двостадійний відбір ще називають *двостадійним відбором елементів* (ДВЕ) (англ. `two-stage element sampling`). Якщо всі вторинні вибіркові одиниці – це кластери елементів, то такий двостадійний відбір називають *двостадійним кластерним відбором* (ДКВ) (англ. `two-stage cluster sampling`).

Якщо процедура відбору складається з трьох або більше стадій, то такий відбір називається *багатостадійним*. У цьому випадку існує чітка ієрархія вибіркових одиниць: генеральна сукупність розглядається як множина первинних вибіркових одиниць, які містять вторинні вибіркові одиниці, а вторинні вибіркові одиниці, в свою чергу, містять третинні вибіркові одиниці, і т. д. Вибіркові одиниці останньої стадії відбору так і називають – *останні вибіркові одиниці*. Якщо всі останні вибіркові одиниці – це окремі елементи, то такий відбір ще називають *багатостадійним відбором елементів* (англ. `multistage element sampling`). Якщо ж останні вибіркові одиниці – це кластери елементів, то такий відбір називають *багатостадійним кластерним відбором* (англ. `multistage cluster sampling`) [20].

8.2. Одностадійний кластерний відбір

8.2.1. Загальний випадок

У випадку одностадійного кластерного відбору скінченна генеральна сукупність $U = \{1, 2, ..., N\}$ ділиться на N_1 підмножин (кластерів):

$$U_1, U_2, ..., U_{N_1}.$$

Множину кластерів символічно зобразимо так:

$$U_{\mathrm{I}} = \{1, 2, \ldots N_{\mathrm{I}}\},$$

тобто, поставимо у відповідність кожному кластеру його номер.

Позначимо кількість елементів у кластері U_i через N_i, $i = 1, \ldots, N_{\mathrm{I}}$. Тоді

$$U = \bigcup_{i \in U_{\mathrm{I}}} U_i; \quad N = \sum_{i \in U_{\mathrm{I}}} N_i.$$

Одностадійний кластерний відбір проводиться так.

1) Із множини кластерів U_{I} за допомогою деякого вибіркового дизайну $p_{\mathrm{I}}(\cdot)$ відбирається ймовірнісна вибірка кластерів s_{I}. Розмір вибірки s_{I} будемо позначати через n_{I} у випадку фіксованого розміру вибірки і через $n_{s_{\mathrm{I}}}$, якщо вибірковий дизайн передбачає змінний розмір вибірки.

2) Обстежуються всі елементи відібраних кластерів.

Вибірковий дизайн $p_{\mathrm{I}}(\cdot)$ може бути будь-яким: простим випадковим без повернення, систематичним, стратифікованим, і т. д.

Якщо позначити через s множину тих елементів, які будуть обстежені, то $s = \bigcup_{i \in s_{\mathrm{I}}} U_i$. Розмір вибірки s дорівнює $n_s = \sum_{s_{\mathrm{I}}} N_i$.

Зауважимо, що навіть якщо $p_{\mathrm{I}}(\cdot)$ є вибірковим дизайном із фіксованим розміром вибірки кластерів, то розмір вибірки елементів не обов'язково фіксований: якщо повторно провести відбір, то при тому самому розмірі нової вибірки кластерів до неї можуть потрапити зовсім інші кластери, розміри яких відрізняються від розмірів тих кластерів, що були в попередній вибірці.

Для вибіркового дизайну $p_{\mathrm{I}}(\cdot)$ ймовірність включення першого порядку для i-го кластера

$$\pi_{\mathrm{I}i} = \sum_{s_{\mathrm{I}} \ni i} p_{\mathrm{I}}(s_{\mathrm{I}}), \quad i = 1, \ldots, N_{\mathrm{I}}.$$

Для двох кластерів i та j ймовірність включення другого порядку

$$\pi_{\mathrm{I}ij} = \sum_{s_{\mathrm{I}} \ni i \& j} p_{\mathrm{I}}(s_{\mathrm{I}}), \quad i, j = 1, \ldots, N_{\mathrm{I}}.$$

Нагадаємо, що $\pi_{\mathrm{I}ii} = \pi_{\mathrm{I}i}$.

Тепер обчислимо ймовірності включення елементів у вибірку. Оскільки вибірка s містить всі елементи відібраних кластерів, то для довільного елемента k з кластера U_i будемо мати

$$\pi_k = P(k \in s) = P(i \in s_{\mathrm{I}}) = \pi_{\mathrm{I}i}, \tag{8.1}$$

де $k = 1, \ldots, N_i$ та $i = 1, \ldots, N_{\mathrm{I}}$.

Ймовірність включення другого порядку обчислюється так:

$$\pi_{kl} = P(k \& l \in s) = \pi_{\mathrm{I}i}, \tag{8.2}$$

якщо елементи k та l обидва належать одному кластеру U_i, та

$$\pi_{kl} = P(k \& l \in s) = P(i \& j \in s_{\mathrm{I}}) = \pi_{\mathrm{I}ij}, \tag{8.3}$$

якщо елементи k та l належать різним кластерам: $k \in U_i$, $l \in U_j$, $i, j = 1, \ldots, N_{\mathrm{I}}$. Нагадаємо, що $\pi_{kk} = \pi_k$.

Нехай $T_i = \sum_{k \in U_i} y_k$ – сумарне значення досліджуваної характеристики в i-му кластері. Тоді $T = \sum_{k \in U} y_k = \sum_{i \in U_{\mathrm{I}}} T_i$ – сумарне значення досліджуваної характеристики для всієї генеральної сукупності.

***Твердження* 8.1.** *При одностадійному кластерному відборі π-оцінка сумарного значення генеральної сукупності має вигляд*

$$\widehat{t}_\pi = \sum_{i \in s_{\mathrm{I}}} \frac{T_i}{\pi_{\mathrm{I}i}}. \tag{8.4}$$

Дисперсія цієї оцінки дорівнює

$$\mathcal{D}(\widehat{t}_\pi) = \sum_{i \in U_{\mathrm{I}}} \sum_{j \in U_{\mathrm{I}}} (\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}) \frac{T_i}{\pi_{\mathrm{I}i}} \frac{T_j}{\pi_{\mathrm{I}j}}. \tag{8.5}$$

Незміщена оцінка дисперсії обчислюється за формулою

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = \sum_{i \in s_{\mathrm{I}}} \sum_{j \in s_{\mathrm{I}}} \frac{\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}}{\pi_{\mathrm{I}ij}} \frac{T_i}{\pi_{\mathrm{I}i}} \frac{T_j}{\pi_{\mathrm{I}j}}. \tag{8.6}$$

□

Якщо $p_{\mathrm{I}}(\cdot)$ – вибірковий дизайн з фіксованим розміром вибірки, то для обчислення дисперсії $\mathcal{D}(\widehat{t}_\pi)$ можна також скористатися формулою Єйтса–Гранді–Сена:

$$\mathcal{D}(\widehat{t}_\pi) = -\frac{1}{2}\sum_{i\in U_{\mathrm{I}}}\sum_{j\in U_{\mathrm{I}}}(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})\left(\frac{T_i}{\pi_{\mathrm{I}i}} - \frac{T_j}{\pi_{\mathrm{I}j}}\right)^2. \qquad (8.7)$$

Незміщену оцінку цієї дисперсії можна обчислити так:

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = -\frac{1}{2}\sum_{i\in s_{\mathrm{I}}}\sum_{j\in s_{\mathrm{I}}}(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})\left(\frac{T_i}{\pi_{\mathrm{I}i}} - \frac{T_j}{\pi_{\mathrm{I}j}}\right)^2. \qquad (8.8)$$

За умови фіксованого розміру вибірки з твердження 8.1 робимо висновки щодо ефективності одностадійного кластерного відбору:

1) Якщо $\forall i = \overline{1, N_{\mathrm{I}}} : \frac{T_i}{\pi_{\mathrm{I}i}} = \mathrm{const}$, то $\mathcal{D}(\widehat{t}_\pi)$=0. Тобто, якщо можна вибрати ймовірності включення $\pi_{\mathrm{I}i}$ так, щоб вони були майже пропорційними до сумарних значень T_i, то дисперсія $\mathcal{D}(\widehat{t}_\pi)$ буде малою і такий кластерний відбір буде дуже ефективним.

2) Якщо розміри кластерів N_i відомі на стадії планування, то ймовірності $\pi_{\mathrm{I}i}$ можна вибрати пропорційними до N_i. Оскільки $T_i = N_i\overline{Y}_i = \sum_{k\in U_i} y_k$, то за умови, що середні значення в кластерах $\overline{Y}_i$ не дуже відрізняються, такий відбір буде дуже ефективним. Якщо ж $\forall i = \overline{1, N_{\mathrm{I}}} : \overline{Y}_i = \mathrm{const}$, то $\mathcal{D}(\widehat{t}_\pi)$=0.

3) Якщо розміри кластерів сильно відрізняються, то ефективність відбору з однаковими ймовірностями (тобто $\pi_{\mathrm{I}i} = \mathrm{const}$) буде низькою. Для того, щоб такий відбір був ефективним, потрібно, щоб $\overline{Y}_i$ були строго пропорційними до N_i^{-1}. У реальних ситуаціях це зустрічається рідко.

8.2.2. Простий випадковий одностадійний кластерний відбір

Розглянемо *простий випадковий (без повернення) одностадійний кластерний відбір* (ПВОКВ). Це означає, що вибірка кластерів s_{I} фіксованого розміру n_{I} вибирається з множини U_{I}, яка містить N_{I} кластерів, за допомогою простого випадкового відбору, після чого обстежуються всі елементи відібраних кластерів. Тоді з розділу 2 та твердження 8.1 будемо мати

$$\widehat{t}_\pi = N_{\mathrm{I}}\overline{T}_{s_{\mathrm{I}}}, \qquad (8.9)$$

де $\overline{T}_{s_{\mathrm{I}}} = \frac{1}{n_{\mathrm{I}}}\sum_{i\in s_{\mathrm{I}}} T_i$ – середнє арифметичне сумарних значень досліджуваної характеристики, обчислене за тими кластерами, які потрапили до вибірки s_{I}.

Дисперсія оцінки сумарного значення обчислюється за формулою

$$\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi) = N_{\mathrm{I}}^2\frac{1-f_{\mathrm{I}}}{n_{\mathrm{I}}}S^2_{TU_{\mathrm{I}}}, \qquad (8.10)$$

де $f_{\mathrm{I}} = \frac{n_{\mathrm{I}}}{N_{\mathrm{I}}}$ – частка відбору кластерів,

$S^2_{TU_{\mathrm{I}}} = \frac{1}{N_{\mathrm{I}}-1}\sum_{i\in U_{\mathrm{I}}}(T_i - \overline{T}_{U_{\mathrm{I}}})^2$ – дисперсія сумарних значень у кластерах,

$\overline{T}_{U_{\mathrm{I}}} = \frac{1}{N_{\mathrm{I}}}\sum_{i\in U_{\mathrm{I}}} T_i$ – середнє арифметичне сумарних значень досліджуваної характеристики, обчислене за всіма кластерами генеральної сукупності.

Незміщена оцінка дисперсії $\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi)$ дорівнює

$$\widehat{\mathcal{D}}_{\text{ПВОКВ}}(\widehat{t}_\pi) = N_{\mathrm{I}}^2\frac{1-f_{\mathrm{I}}}{n_{\mathrm{I}}}S^2_{Ts_{\mathrm{I}}}, \qquad (8.11)$$

де $S^2_{Ts_{\mathrm{I}}} = \frac{1}{n_{\mathrm{I}}-1}\sum_{i\in s_{\mathrm{I}}}(T_i - \overline{T}_{s_{\mathrm{I}}})^2$.

Зауваження 8.1. Систематичний випадковий відбір з одним випадковим стартом є простим випадковим одностадійним кластерним відбором з $n_{\mathrm{I}} = 1$, при цьому кількість можливих систематичних вибірок $a = N_{\mathrm{I}}$. Систематичний відбір з m випадковими стартами можна розглядати як ПВОКВ з $n_{\mathrm{I}} = m$ та $N_{\mathrm{I}} = ma$.

Розглянемо *коефіцієнт однорідності* $\delta = 1 - \frac{S_W^2}{S^2}$, де $S_W^2 = \frac{1}{N-N_\text{I}} \sum\limits_{i\in U_\text{I}} \sum\limits_{k\in U_i} (y_k - \overline{Y}_i)^2$ – сукупна внутрішньокластерна дисперсія, $\overline{Y}_i = \sum\limits_{k\in U_i} \frac{y_k}{N_i}$ – середнє в i-му кластері, $i \in U_\text{I}$.

З іншого боку дисперсію S_W^2 можна записати так:

$$S_W^2 = \frac{\sum\limits_{i\in U_I} (N_i - 1)S_i^2}{\sum\limits_{i\in U_I} (N_i - 1)},$$

де $S_i^2 = \frac{1}{N_i-1} \sum\limits_{k\in U_i} (y_k - \overline{Y}_i)^2$ – це дисперсія в кластері U_i, тобто S_W^2 – це зважене середнє N_I дисперсій S_i^2 з вагами $N_i - 1$.

Отже, коефіцієнт δ додатний або від'ємний залежно від того, більша чи менша зважена внутрішньокластерна дисперсія порівняно із загальною дисперсією в генеральній сукупності.

Коефіцієнт однорідності δ має такі обмеження:

$$-\frac{N_\text{I} - 1}{N - N_\text{I}} \le \delta \le 1.$$

Якщо $\delta = 1$, то варіація всередині кластерів рівна 0. Якщо $\delta = -\frac{N_\text{I}-1}{N-N_\text{I}}$, то середні у всіх кластерах однакові. Нижня межа $\left(-\frac{N_\text{I}-1}{N-N_\text{I}}\right)$ зазвичай близька до нуля, особливо якщо N велике порівняно з N_I. Значення $\delta = 0$ досягається тоді, коли зважена внітрішньокластерна дисперсія дорівнює загальній дисперсії в генеральній сукупності. Малі значення δ означають, що елементи всередині кластерів неоднорідні відносно досліджуваної характеристики. І навпаки, великі значення δ свідчать, що елементи всередині кластерів однорідні відносно досліджуваної характеристики.

Позначимо через $\overline{N} = N/N_\text{I}$ середню кількість елементів на один кластер та запишемо коваріацію між N_i та $N_i\overline{Y}_i^2$:

$$\text{Cov} = \frac{1}{N_\text{I} - 1} \sum_{i\in U_i} (N_i - \overline{N}) N_i \overline{Y}_i^2.$$

Легко перевірити, що

$$S_{TU_\text{I}}^2 = \overline{N} S^2 \left(1 + \frac{N - N_\text{I}}{N_\text{I} - 1}\delta\right) + \text{Cov}.$$

Якщо підставити цей вираз у формулу (8.10), то отримаємо такий вигляд дисперсії $\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi)$:

$$\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi) = \left(1 + \frac{N - N_\text{I}}{N_\text{I} - 1}\delta\right) \overline{N} K_\text{I} S^2 + K_\text{I}\text{Cov},$$

де $K_\text{I} = N_\text{I}^2(1 - f_\text{I})/n_\text{I}$.

Математичне сподівання кількості обстежених елементів при ПВОКВ з розміром вибірки кластерів n_I обчислюється так:

$$E_{\text{ПВОКВ}}(n_s) = n_\text{I}\overline{N} =: n.$$

Для об'єктивного порівняння простого випадкового відбору з простим випадковим одностадійним кластерним відбором припустимо, що розмір простої випадкової вибірки $n = n_\text{I}\overline{N}$. При ПВВ π-оцінка сумарного значення дорівнює $N\overline{y}$, а дисперсія цієї оцінки

$$\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi) = \mathcal{D}_{\text{ПВВбП}}(N\overline{y}) = \overline{N} K_\text{I} S^2.$$

Звідси отримуємо ще одну формулу для обчислення дисперсії оцінки $\widehat{t}_\pi$ при ПВОКВ:

$$\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi) = \left(1 + \frac{N - N_\text{I}}{N_\text{I} - 1}\delta\right) \mathcal{D}_{\text{ПВВбП}} + K_\text{I}\text{Cov}. \qquad (8.12)$$

Отже, дизайн-ефект простого випадкового одностадійного кластерного відбору дорівнює

$$deff(\text{ПВОКВ}, \widehat{t}_\pi) = \frac{\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi)}{\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)} = 1 + \frac{N - N_\text{I}}{N_\text{I} - 1}\delta + \frac{\text{Cov}}{\overline{N} S^2}. \qquad (8.13)$$

Тепер дослідимо, наскільки ефективним є простий випадковий одностадійний кластерний відбір. Для цього розглянемо два випадки.

1) *Кластери однакового розміру.* Припустимо, що для всіх $i \in U_1 : N_i = \overline{N}$. Тоді $\mathrm{Cov} = 0$ та

$$deff(\text{ПВОКВ}, \widehat{t}_\pi) = 1 + \frac{N - N_1}{N_1 - 1}\delta \approx 1 + (\overline{N} - 1)\delta. \qquad (8.14)$$

Це означає, що $\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi) < \mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)$ тоді і тільки тоді, коли $\delta < 0$, тобто, тоді і тільки тоді, коли внутрішньокластерна дисперсія досить велика. Але на практиці більшість кластерів утворюють за принципом сусідніх елементів, а такі елементи зазвичай мають схожі характеристики, тому більш імовірно, що $\delta > 0$. Отже, зазвичай $\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi)$ перевищує $\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)$. Наприклад, при досить малому додатному $\delta = 0,08$ та середньому розмірі кластерів $\overline{N} = 300$ будемо мати

$$deff(\text{ПВОКВ}, \widehat{t}_\pi) \approx 25.$$

У даному випадку застосування кластерного відбору з досить великими розмірами кластерів призводить до значної втрати ефективності.

2) *Кластери різних розмірів.* Якщо кореляція між N_i та $N_i\overline{Y}_i^2$ додатна, як зазвичай буває, то зростання дисперсії внаслідок використання кластерного відбору може бути навіть більшим, ніж у попередньому випадку, оскільки другий доданок у виразі (8.12) може бути більшим. Покажемо, наскільки важливим є те, що розміри кластерів відрізняються. Для цього розглянемо граничний випадок $\delta = \delta_{\min} = (N_1 - 1)/(N - N_1)$. Тут середні значення однакові у всіх кластерах і дорівнюють $\overline{Y}$. Тоді формула (8.12) матиме вигляд

$$\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi) = \overline{Y}^2 K_1 S^2_{NU_1}. \qquad (8.15)$$

А значення цієї дисперсії буде тим більшим, чим більше значення дисперсії розмірів кластерів $S^2_{NU_1} = \frac{1}{N_1 - 1}\sum\limits_{i \in U_1}(N_i - \overline{N})^2$.

У цьому випадку

$$deff(\text{ПВОКВ}, \widehat{t}_\pi) = \overline{N}\left(\frac{CV_N}{CV_y}\right)^2, \qquad (8.16)$$

де $CV_N = S_{NU_1}/\overline{N}$ та $CV_y = S/\overline{Y}$. Відношення $\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi)/\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)$ може бути значно більшим від одиниці, особливо тоді, коли середній розмір кластерів $\overline{N}$ великий.

Отже, одностадійний кластерний відбір з наступним використанням π-оцінки для оцінювання сумарного значення досліджуваної характеристики в багатьох ситуаціях може бути неефективним, особливо якщо кластери однорідні та/чи неоднакового розміру. Але з точки зору вартості обстеження така стратегія може мати переваги.

Якщо наявна додаткова інформація, то ефективність такого відбору можна підвищити. Тоді вибір стратегії залежить від того, яка саме інформація є у розпорядженні дослідника. Це може бути як відбір кластерів з імовірностями, пропорційними деякій змінній розміру, так і використання іншого методу оцінювання.

8.3. Двостадійний відбір

З вищенаведеного можна зробити висновок, що ефективність кластерного відбору зазвичай менша, ніж ефективність простого випадкового відбору. Причиною цього є те, що елементи всередині кластерів, як правило, досить однорідні ($\delta > 0$), та й кластери частіше бувають різні за розміром.

Дисперсію π-оцінки можна зменшити за рахунок збільшення кількості кластерів у вибірці. Але це може призвести до збільшення витрат при обмеженому бюджеті.

Для того, щоб укластися в рамки бюджету і одночасно збільшити кількість кластерів у вибірці, можна робити підвибірки з відібраних кластерів, а не обстежувати всі елементи. Якщо елементи в кластерах однорідні, тобто варіація всередині кластера мала, то оцінки $\widehat{t}_i$ сумарних значень в кластерах будуть мати малу дисперсію навіть при досить малих підвибірках. Тому дуже часто замість одностадійного кластерного відбору використовують двостадійний відбір.

У випадку двостадійного відбору існує два джерела вибіркових похибок (тобто, похибок внаслідок обстеження частини генераль-

ної сукупності замість суцільного обстеження): на першій стадії відбору – це похибка внаслідок вибору первинних вибіркових одиниць (ПВО), а на другій стадії – похибка через відбір вторинних вибіркових одиниць(ВВО) з тих ПВО, що потрапили до вибірки на попередній стадії.

Генеральна сукупність $U = \{1, 2, ..., k, ..., N\}$ поділена на N_{I} первинних вибіркових одиниць $U_1, U_2, ..., U_{N_{\text{I}}}$. Позначимо множину (популяцію) первинних вибіркових одиниць через $U_{\text{I}} = \{1, 2, ..., i, ..., N_{\text{I}}\}$. Розмір первинної вибіркової одиниці U_i, тобто кількість елементів в ній, позначимо через N_i. Матимемо

$$N = \sum_{i \in U_{\text{I}}} N_i.$$

Опишемо загальну процедуру двостадійного відбору.

Перша стадія. Відбирається вибірка s_{I} первинних вибіркових одиниць із множини U_{I} ($s_{\text{I}} \subset U_{\text{I}}$) згідно з вибірковим дизайном $p_{\text{I}}(\cdot)$.

Друга стадія. Для кожного $i \in s_{\text{I}}$ відбирається вибірка елементів, яку ми позначимо s_i, з первинної вибіркової одиниці U_i ($s_i \subset U_i$) згідно з вибірковим дизайном $p_i(\cdot|s_{\text{I}})$. Результуюча вибірка елементів, яку позначимо як s, є об'єднанням вибірок з відібраних на першій стадії первинних вибіркових одиниць: $s = \bigcup_{i \in s_{\text{I}}} s_i$.

Як на першій, так і на другій стадії відбору можна використовувати будь-який вибірковий дизайн. Причому вибір методів відбору, які планується застосувати на другій стадії, може залежати від результату s_{I} першої стадії відбору. Більше того, відбір з ПВО U_i може залежати від відбору з ПВО U_j, $i \neq j$.

Але в даному розділі ми зосередимо увагу не на загальному випадку, а розглянемо вужчий клас вибіркових дизайнів, які можна застосовувати на другій стадії відбору, а саме: будемо вимагати, щоб виконувалися умови *інваріантності* та *незалежності*.

Інваріантність вибіркового дизайну другої стадії відбору означає, що $\forall\, i \in U_{\text{I}}$ та $\forall\, s_{\text{I}} \ni i$ виконується умова $p_i(\cdot|s_{\text{I}}) = p_i(\cdot)$. Інакше кажучи, кожен раз, коли i-та ПВО потрапляє у вибірку

першої стадії відбору, на другій стадії для відбору з цієї одиниці потрібно використовувати один і той самий вибірковий дизайн (незалежно від того, які ще ПВО потрапили у вибірку).

Незалежність вибіркового дизайну другої стадії відбору означає, що $\forall\, s_{\text{I}}$

$$P\left(\bigcup_{i \in s_{\text{I}}} s_i \,\middle|\, s_{\text{I}}\right) = \prod_{i \in s_{\text{I}}} P\left(s_i | s_{\text{I}}\right),$$

тобто, відбір з кожної ПВО не залежить від відборів з інших ПВО.

8.3.1. Двостадійний відбір елементів

Надалі будемо припускати, що умови інваріантності та незалежності виконуються, і сконцентруємо увагу на випадку двостадійного відбору елементів. Це означає, що вторинні вибіркові одиниці – окремі елементи.

Кількість ПВО у вибірці s_{I} позначимо через $n_{s_{\text{I}}}$ або, якщо $p_{\text{I}}(\cdot)$ – вибірковий дизайн з фіксованим розміром вибірки, то n_{I}. Кількість елементів у вибірці s_i позначимо через n_{s_i} або, якщо $p_i(\cdot)$ – вибірковий дизайн з фіксованим розміром вибірки, то n_i. Тоді загальна кількість елементів у вибірці s дорівнює $n_s := \sum_{i \in s_{\text{I}}} n_{s_i}$.

Обчислимо ймовірність включення елемента у вибірку за двостадійного відбору. Для вибіркового дизайну $p_{\text{I}}(\cdot)$ першої стадії відбору ймовірності включення ПВО у вибірку позначимо через $\pi_{\text{I}i}$ та $\pi_{\text{I}ij}$. Для вибіркового дизайну $p_i(\cdot)$ другої стадії відбору ймовірності включення ВВО у вибірку будемо позначати через $\pi_{k|i}$ та $\pi_{kl|i}$. Тоді ймовірність включення k-го елемента генеральної сукупності у вибірку обчислюється так:

$$\pi_k = \pi_{\text{I}i}\pi_{k|i}, \quad \text{де} \quad i \in U_{\text{I}} : U_i \ni k, \tag{8.17}$$

а ймовірність того, що k-й та l-й елементи генеральної сукупності потраплять до вибірки, дорівнює

$$\pi_{kl} = \begin{cases} \pi_{\text{I}i}\pi_{k|i}, & \text{якщо} \quad k = l \in U_i; \\ \pi_{\text{I}i}\pi_{kl|i}, & \text{якщо} \quad k \& l \in U_i, k \neq l; \\ \pi_{\text{I}ij}\pi_{k|j}\pi_{l|j}, & \text{якщо} \quad k \in U_i, l \in U_j, (i \neq j). \end{cases} \tag{8.18}$$

Оцінка Горвіца–Томпсона сумарного значення $T_i = \sum\limits_{k \in U_i} y_k$ для первинної вибіркової одиниці U_i, обчислена за «другостадійною» вибіркою з цієї одиниці, знаходиться так:

$$\widehat{t}_{i\pi} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}}. \tag{8.19}$$

Дисперсія цієї оцінки обчислюється за формулою

$$\mathcal{D}_i = \sum_{k \in U_i} \sum_{l \in U_i} \left(\pi_{kl|i} - \pi_{k|i}\pi_{l|i}\right) \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}, \tag{8.20}$$

а її незміщена оцінка дорівнює

$$\widehat{\mathcal{D}}_i = \sum_{k \in s_i} \sum_{l \in s_i} \frac{\left(\pi_{kl|i} - \pi_{k|i}\pi_{l|i}\right)}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}. \tag{8.21}$$

***Твердження* 8.2.** *У випадку двостадійного відбору елементів π-оцінка сумарного значення $T = \sum\limits_{k \in U} y_k$ досліджуваної характеристики має вигляд*

$$\widehat{t}_\pi = \sum_{i \in s_\mathrm{I}} \frac{\widehat{t}_{i\pi}}{\pi_{\mathrm{I}i}}, \tag{8.22}$$

де $\widehat{t}_{i\pi}$ – π-оцінка сумарного значення T_i, $i \in s_\mathrm{I}$.

Дисперсія оцінки $\widehat{t}_\pi$ має дві складові:

$$\mathcal{D}\left(\widehat{t}_\pi\right) = \mathcal{D}_{\mathrm{ПВО}} + \mathcal{D}_{\mathrm{ВВО}}, \tag{8.23}$$

де

$$\mathcal{D}_{\mathrm{ПВО}} = \sum_{i \in U_\mathrm{I}} \sum_{j \in U_\mathrm{I}} \left(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}\right) \frac{T_i}{\pi_{\mathrm{I}i}} \frac{T_j}{\pi_{\mathrm{I}j}}, \tag{8.24}$$

$$\mathcal{D}_{\mathrm{ВВО}} = \sum_{i \in U_\mathrm{I}} \frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}}. \tag{8.25}$$

Незміщеною оцінкою дисперсії $\mathcal{D}_{\mathrm{ПВО}}$ є статистика

$$\begin{aligned} \widehat{\mathcal{D}}_{\mathrm{ПВО}} &= \sum_{i \in s_\mathrm{I}} \sum_{j \in s_\mathrm{I}} \frac{\left(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}\right)}{\pi_{\mathrm{I}ij}} \frac{\widehat{t}_{i\pi}}{\pi_{\mathrm{I}i}} \frac{\widehat{t}_{j\pi}}{\pi_{\mathrm{I}j}} - \\ &- \sum_{i \in s_\mathrm{I}} \frac{1}{\pi_{\mathrm{I}i}} \left(\frac{1}{\pi_{\mathrm{I}i}} - 1\right) \widehat{\mathcal{D}}_i, \end{aligned} \tag{8.26}$$

а незміщена оцінка дисперсії $\mathcal{D}_{\mathrm{ВВО}}$ дорівнює

$$\widehat{\mathcal{D}}_{\mathrm{ВВО}} = \sum_{i \in s_\mathrm{I}} \frac{\widehat{\mathcal{D}}_i}{\left(\pi_{\mathrm{I}i}\right)^2}. \tag{8.27}$$

Отже, незміщена оцінка дисперсії $\mathcal{D}\left(\widehat{t}_\pi\right)$ має вигляд

$$\begin{aligned} \widehat{\mathcal{D}}\left(\widehat{t}_\pi\right) &= \widehat{\mathcal{D}}_{\mathrm{ПВО}} + \widehat{\mathcal{D}}_{\mathrm{ВВО}} = \\ &= \sum_{i \in s_\mathrm{I}} \sum_{j \in s_\mathrm{I}} \frac{\left(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}\right)}{\pi_{\mathrm{I}ij}} \frac{\widehat{t}_{i\pi}}{\pi_{\mathrm{I}i}} \frac{\widehat{t}_{j\pi}}{\pi_{\mathrm{I}j}} + \sum_{i \in s_\mathrm{I}} \frac{\widehat{\mathcal{D}}_i}{\left(\pi_{\mathrm{I}i}\right)}. \end{aligned} \tag{8.28}$$

Доведення. З означення π-оцінки та формул (8.17) і (8.19) отримуємо формулу для обчислення π-оцінки сумарного значення у випадку двостадійного відбору елементів:

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{i \in s_\mathrm{I}} \sum_{k \in s_i} \frac{y_k}{\pi_{\mathrm{I}i}\pi_{k|i}} = \sum_{i \in s_\mathrm{I}} \frac{1}{\pi_{\mathrm{I}i}} \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \sum_{i \in s_\mathrm{I}} \frac{\widehat{t}_{i\pi}}{\pi_{\mathrm{I}i}}.$$

Щоб отримати формулу (8.23) для обчислення дисперсії π-оцінки сумарного значення, скористаємося властивостями умовного математичного сподівання:

$$\mathcal{D}(\widehat{t}_\pi) = \mathcal{D}_{p_\mathrm{I}}[E(\widehat{t}_\pi|s_\mathrm{I})] + E_{p_\mathrm{I}}[\mathcal{D}(\widehat{t}_\pi|s_\mathrm{I})]. \tag{8.29}$$

Розглянемо $E(\widehat{t}_\pi|s_\mathrm{I})$ та $\mathcal{D}(\widehat{t}_\pi|s_\mathrm{I})$ окремо і скористаємося властивостями інваріантності та незалежності вибіркових дизайнів $(p_i, i \in U_\mathrm{I})$ другої стадії відбору:

$$E(\widehat{t}_\pi|s_\mathrm{I}) = \sum_{i \in s_\mathrm{I}} E_{p_i}\left(\frac{\hat{t}_{i\pi}}{\pi_{\mathrm{I}i}}\middle| s_\mathrm{I}\right) = \sum_{i \in s_\mathrm{I}} E_{p_i}\left(\frac{\hat{t}_{i\pi}}{\pi_{\mathrm{I}i}}\right) = \sum_{i \in s_\mathrm{I}} \frac{T_i}{\pi_{\mathrm{I}i}}, \tag{8.30}$$

$$\mathcal{D}(\widehat{t}_\pi|s_\mathrm{I}) = \sum_{i \in s_\mathrm{I}} \mathcal{D}\left(\frac{\hat{t}_{i\pi}}{\pi_{\mathrm{I}i}}\middle| s_\mathrm{I}\right) = \sum_{i \in s_\mathrm{I}} \mathcal{D}_{p_i}\left(\frac{\hat{t}_{i\pi}}{\pi_{\mathrm{I}i}}\right) = \sum_{i \in s_\mathrm{I}} \frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}^2}. \tag{8.31}$$

Із виразів (8.29)–(8.31) отримуємо

$$\mathcal{D}(\widehat{t}_\pi) = \mathcal{D}_{p_\mathrm{I}}\left[\sum_{i\in s_\mathrm{I}}\frac{T_i}{\pi_{\mathrm{I}i}}\right] + E_{p_\mathrm{I}}\left[\sum_{i\in s_\mathrm{I}}\frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}^2}\right] =$$

$$= \sum_{i\in U_\mathrm{I}}\sum_{j\in U_\mathrm{I}}(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})\frac{T_i}{\pi_{\mathrm{I}i}}\frac{T_j}{\pi_{\mathrm{I}j}} + \sum_{i\in U_\mathrm{I}}\frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}},$$

тобто, маємо вираз (8.23), що й треба було довести.

Тепер розглянемо компоненти оцінки дисперсії. Внаслідок властивості незалежності маємо

$$E(\hat{t}_{i\pi}\hat{t}_{j\pi}|s_\mathrm{I}) = \begin{cases} T_i^2 + \mathcal{D}_i, & i = j; \\ T_iT_j, & i \neq j. \end{cases}$$

Обчислимо математичне сподівання першого доданку у виразі (8.26), тобто:

$$E\left(\sum_{i\in s_\mathrm{I}}\sum_{j\in s_\mathrm{I}}\frac{(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})}{\pi_{\mathrm{I}ij}}\frac{\widehat{t}_{i\pi}}{\pi_{\mathrm{I}i}}\frac{\widehat{t}_{j\pi}}{\pi_{\mathrm{I}j}}\right) =$$

$$= E_{p_\mathrm{I}}\left[\sum_{i\in s_\mathrm{I}}\sum_{j\in s_\mathrm{I}}\frac{(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})}{\pi_{\mathrm{I}ij}}\frac{E(\widehat{t}_{i\pi}\widehat{t}_{j\pi}|s_\mathrm{I})}{\pi_{\mathrm{I}i}\pi_{\mathrm{I}j}}\right] =$$

$$= E_{p_\mathrm{I}}\left(\sum_{i\in s_\mathrm{I}}\sum_{j\in s_\mathrm{I}}\frac{(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})}{\pi_{\mathrm{I}ij}}\frac{T_iT_j}{\pi_{\mathrm{I}i}\pi_{\mathrm{I}j}}\right) +$$

$$+E_{p_\mathrm{I}}\left(\sum_{i\in s_\mathrm{I}}\frac{\pi_{\mathrm{I}i}(1-\pi_{\mathrm{I}i})}{\pi_{\mathrm{I}i}}\frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}^2}\right) =$$

$$= \sum_{i\in U_\mathrm{I}}\sum_{j\in U_\mathrm{I}}(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})\frac{T_iT_j}{\pi_{\mathrm{I}i}\pi_{\mathrm{I}j}} + \sum_{i\in U_\mathrm{I}}\left(\frac{1}{\pi_{\mathrm{I}i}} - 1\right)\mathcal{D}_i =$$

$$= \mathcal{D}_{\mathrm{ПВО}} + \sum_{i\in U_\mathrm{I}}\left(\frac{1}{\pi_{\mathrm{I}i}} - 1\right)\mathcal{D}_i.$$

Для математичного сподівання другого доданка у виразі (8.26) будемо мати

$$E\left[-\sum_{i\in s_\mathrm{I}}\frac{1}{\pi_{\mathrm{I}i}}\left(\frac{1}{\pi_{\mathrm{I}i}} - 1\right)\widehat{\mathcal{D}}_i\right] =$$

$$= -E_{p_\mathrm{I}}\left[\sum_{i\in s_\mathrm{I}}\frac{1}{\pi_{\mathrm{I}i}}\left(\frac{1}{\pi_{\mathrm{I}i}} - 1\right)E(\widehat{\mathcal{D}}_i|s_\mathrm{I})\right] =$$

$$= -E_{p_\mathrm{I}}\left[\sum_{i\in s_\mathrm{I}}\frac{1}{\pi_{\mathrm{I}i}}\left(\frac{1}{\pi_{\mathrm{I}i}} - 1\right)\mathcal{D}_i\right] =$$

$$= -\sum_{i\in U_\mathrm{I}}\left(\frac{1}{\pi_{\mathrm{I}i}} - 1\right)\mathcal{D}_i.$$

Звідси випливає, що $E(\widehat{\mathcal{D}}_{\mathrm{ПВО}}) = \mathcal{D}_{\mathrm{ПВО}}$.

Аналогічно

$$E(\widehat{\mathcal{D}}_{\mathrm{ВВО}}) = E_{p_\mathrm{I}}\left[\sum_{i\in s_\mathrm{I}}\frac{E(\widehat{\mathcal{D}}_i|s_\mathrm{I})}{\pi_{\mathrm{I}i}^2}\right] =$$

$$= E_{p_\mathrm{I}}\left[\sum_{i\in s_\mathrm{I}}\frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}^2}\right] = \sum_{i\in U_\mathrm{I}}\frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}} = \mathcal{D}_{\mathrm{ВВО}}.$$

Отже,

$$E(\widehat{\mathcal{D}}(\widehat{t}_\pi)) = E(\widehat{\mathcal{D}}_{\mathrm{ПВО}} + \widehat{\mathcal{D}}_{\mathrm{ВВО}}) = \mathcal{D}_{\mathrm{ПВО}} + \mathcal{D}_{\mathrm{ВВО}} = \mathcal{D}(\widehat{t}_\pi),$$

тобто незміщеність оцінки $E(\widehat{\mathcal{D}}(\widehat{t}_\pi))$ доведена.

□

Зауваження 8.2. Оцінка (8.26) не завжди набуває додатних значень.

Наведемо умови, за яких складові дисперсії оцінки $\widehat{t}_\pi$ нульові.

1) Якщо $s_{\mathrm{I}} = U_{\mathrm{I}}$ з імовірністю одиниця, то $\pi_{\mathrm{I}i} = \pi_{\mathrm{I}ij} = 1\ \forall i, j$, а отже $\mathcal{D}_{\mathrm{ПВО}} = 0$ та $\mathcal{D}_{\mathrm{ВВО}} = \sum\limits_{i \in U_{\mathrm{I}}} \mathcal{D}_i$. Саме такий вигляд має дисперсія оцінки $\widehat{t}_\pi$ у випадку стратифікованого відбору, а пояснюється це тим, що страти можна розглядати як первинні вибіркові одиниці двостадійного відбору, в якому $s_{\mathrm{I}} = U_{\mathrm{I}}$.

2) Якщо $s_i = U_i$ з імовірністю одиниця, то $\mathcal{D}_{\mathrm{ВВО}} = 0$. Тоді $\mathcal{D}_{\mathrm{ПВО}}$ – це дисперсія π-оцінки при одностадійному кластерному відборі.

Обчислити оцінку дисперсії $\mathcal{D}\left(\widehat{t}_\pi\right)$ за формулою (8.28) може бути досить складно – потрібно оцінити $\widehat{\mathcal{D}}_i$ для всіх значень $i \in s_{\mathrm{I}}$. Тому інколи $\mathcal{D}\left(\widehat{t}_\pi\right)$ оцінюють за формулою

$$\widehat{\mathcal{D}}^* = \sum_{i \in s_{\mathrm{I}}} \sum_{j \in s_{\mathrm{I}}} \frac{(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})}{\pi_{\mathrm{I}ij}} \frac{\widehat{t}_{i\pi}}{\pi_{\mathrm{I}i}} \frac{\widehat{t}_{j\pi}}{\pi_{\mathrm{I}j}}.$$

Щоб обчислити цю оцінку, потрібно оцінити тільки сумарні значення в ПВО, а це значно простіше, ніж оцінювання дисперсій. Ця оцінка є зміщеною. Її зміщення дорівнює:

$$B\left(\widehat{\mathcal{D}}^*\right) = -\sum_{i \in U_i} \mathcal{D}_i,$$

тобто $\widehat{\mathcal{D}}^*$ недооцінює невідому реальну дисперсію оцінки $\widehat{t}_\pi$.

Проте, якщо розглянути відносне зміщення цієї оцінки:

$$\frac{B\left(\widehat{\mathcal{D}}^*\right)}{\mathcal{D}\left(\widehat{t}_\pi\right)} = -\frac{\sum\limits_{i \in U_{\mathrm{I}}} \mathcal{D}_i}{\sum\limits_{i \in U_{\mathrm{I}}} \sum\limits_{j \in U_{\mathrm{I}}} (\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}) \frac{T_i T_j}{\pi_{\mathrm{I}i}\pi_{\mathrm{I}j}} + \sum\limits_{i \in U_{\mathrm{I}}} \frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}}}, \tag{8.32}$$

то можна побачити, що в багатьох випадках зміщення оцінки $\widehat{D}^*$ може бути несуттєвим. Якщо ймовірності $\pi_{\mathrm{I}i}$ – малі, то і чисельник дробу у виразі (8.32) буде малим порівняно зі знаменником, внаслідок чого зміщення буде незначним і ним можна буде знехтувати.

Приклад 8.1. Нехай відбір кластерів проводиться за допомогою простого випадкового відбору з імовірностями включення $\pi_{\mathrm{I}i} = \frac{n_{\mathrm{I}}}{N_{\mathrm{I}}} = 0,1$. Якщо, наприклад, $\mathcal{D}_{\mathrm{ПВО}}/\mathcal{D}_{\mathrm{ВВО}} = 5$, то відносне зміщення, обчислене за формулою (8.32), дорівнює: $-\frac{1}{60} = -0,017$. ◊

8.3.2. Самозважений двостадійний відбір

У багатьох двостадійних та багатостадійних обстеженнях використовують так званий *самозважений вибірковий дизайн*. Розглянемо самозважений двостадійний відбір.

Нехай u_i – це відома (хоча б приблизно) міра розміру i-ї ПВО, $i \in U_{\mathrm{I}}$. Тоді на першій стадії відбору можна використати дизайн з імовірностями включення $\pi_{\mathrm{I}i} = cu_i$, де c – деяка константа, а на другій стадії провести простий випадковий відбір n_i елементів з N_i так, щоб

$$\frac{n_i}{N_i} = \frac{1}{u_i}.$$

Якщо на першій стадії відбору вибірка має фіксований розмір n_{I}, то $c = \frac{n_{\mathrm{I}}}{\sum\limits_{i \in U_{\mathrm{I}}} u_i}$ та для будь-якого елемента k отримаємо

$$\pi_k = \pi_{\mathrm{I}i}\pi_{k|i} = cu_i \frac{n_i}{N_i} = cu_i \frac{1}{u_i} = c.$$

Тоді π-оцінка сумарного значення T буде мати вигляд

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{1}{c} \sum_{i \in s_{\mathrm{I}}} \sum_{k \in s_i} y_k.$$

Отже, всі значення y_k мають однакову вагу, яка дорівнює $\frac{1}{c}$.

8.3.3. Простий випадковий відбір на обох стадіях двостадійного відбору

Нехай на обох стадіях двостадійного відбору елементів використовується простий випадковий відбір без повернення: на першій стадії відбирається проста випадкова вибірка s_{I} розміру n_{I} з N_{I}

первинних вибіркових одиниць, та на другій стадії для всіх $i \in s_1$ вибираємо n_i елементів (ВВО) з N_i елементів i-ї ПВО. Такий відбір надалі будемо називати *простим випадковим двостадійним відбором елементів* і позначати через ПВДВЕ.

Тоді π-оцінка сумарного значення досліджуваної характеристики генеральної сукупності

$$\widehat{t}_\pi = \frac{N_\text{I}}{n_\text{I}} \sum_{i \in s_1} N_i \overline{y}_i = \frac{N_\text{I}}{n_\text{I}} \sum_{i \in s_\text{I}} \widehat{t}_{i\pi}, \tag{8.33}$$

де

$$\widehat{t}_{i\pi} = N_i \overline{y}_i = \frac{N_i}{n_i} \sum_{k \in s_i} y_k.$$

Дисперсія π-оцінки сумарного значення генеральної сукупності

$$\mathcal{D}_{\text{ПВДВЕ}}\left(\widehat{t}_\pi\right) = {N_\text{I}}^2 \frac{1 - f_\text{I}}{n_\text{I}} S_T^2 + \frac{N_\text{I}}{n_\text{I}} \sum_{i \in U_\text{I}} N_i^2 \frac{1 - f_i}{n_i} S_i^2, \tag{8.34}$$

де

$S_T^2(t) = \frac{1}{N_\text{I}-1} \sum\limits_{i \in U_\text{I}} \left(T_i - \overline{T}\right)^2$ – дисперсія сумарних значень досліджуваної характеристики y в первинних вибіркових одиницях,

$\overline{T} = \frac{1}{N_\text{I}} \sum\limits_{i \in U_\text{I}} T_i$ – середнє арифметичне сумарних значень досліджуваної характеристики y в первинних вибіркових одиницях,

$S_i^2 = \frac{1}{N_i - 1} \sum\limits_{k \in U_i} \left(y_k - \overline{Y}_i\right)^2$ – дисперсія досліджуваної характеристики в i-й ПВО,

$\overline{Y}_i = \frac{1}{N_i} \sum\limits_{k \in U_i} y_k$ – середнє досліджуваної характеристики в i-й ПВО, $i \in U_\text{I}$.

Незміщена оцінка для $\mathcal{D}\left(\widehat{t}_\pi\right)$ має вигляд

$$\widehat{\mathcal{D}}_{\text{ПВДВЕ}}\left(\widehat{t}_\pi\right) = {N_\text{I}}^2 \frac{1 - f_\text{I}}{n_\text{I}} \widehat{S}_T^2 + \frac{N_\text{I}}{n_\text{I}} \sum_{i \in s_\text{I}} N_i^2 \frac{1 - f_i}{n_i} \widehat{S}_i^2, \tag{8.35}$$

де $\widehat{S}_T^2 = \frac{1}{n_\text{I}-1} \sum\limits_{i \in s_\text{I}} \left(\widehat{t}_{i\pi} - \frac{1}{n_\text{I}} \sum\limits_{i \in s_\text{I}} \widehat{t}_{i\pi}\right)^2$ – оцінка дисперсії сумарних значень у ПВО,

$\widehat{t}_{i\pi} = N_i \overline{y}_i$ – оцінка сумарного значення досліджуваної характеристики в i-й ПВО,

$\overline{y}_i = \frac{1}{n_i} \sum\limits_{k \in s_i} y_k$ – вибіркове середнє досліджуваної характеристики в i-й ПВО,

$\widehat{S}_i^2 = \frac{1}{n_\text{I}-1} \sum\limits_{k \in s_\text{I}} \left(y_k - \overline{y}_i\right)^2$ – вибіркова дисперсія досліджуваної характеристики в i-й ПВО.

8.3.4. Оптимальне розміщення у випадку простого випадкового двостадійного відбору елементів

Розглянемо простий випадковий двостадійний відбір елементів, описаний у попередньому параграфі, та спробуємо визначити оптимальну частку відбору $f_\text{I} = n_\text{I}/N_\text{I}$ для вибірки першої стадії, а також частки відборів $f_i = n_i/N_i$ для вибірок другої стадії цього відбору, пам'ятаючи про такі очевидні обмеження:

$$0 < f_i \leq 1, \quad i \in U_\text{I}; \tag{8.36}$$

$$0 < f_\text{I} \leq 1. \tag{8.37}$$

У випадку простого випадкового двостадійного відбору елементів π-оцінка сумарного значення досліджуваної характеристики генеральної сукупності обчислюється згідно з формулою (8.33), а дисперсія цієї оцінки – за формулою (8.34). Цю дисперсію можна переписати у вигляді

$$\mathcal{D}_{\text{ПВДВЕ}}(\widehat{t}_\pi) = A_0 + \frac{A_1}{f_\text{I}} + \frac{1}{f_\text{I}} \sum_{i \in U_i} \frac{A_{2i}}{f_i}, \tag{8.38}$$

де

$$A_0 = -N_\text{I} S_{TU_\text{I}}^2; \quad A_1 = N_\text{I} G; \quad A_{2i} = N_i S_i^2; \tag{8.39}$$

$$G = S_{TU_\text{I}}^2 - \frac{1}{N_\text{I}} \sum_{i \in U_\text{I}} N_i S_i^2. \tag{8.40}$$

Розглянемо функцію витрат виду

$$C(s_\text{I}) = c_0 + n_\text{I} c_u + \left(\sum_{i \in s_\text{I}} N_i\right) c_{el} + \sum_{i \in s_\text{I}} n_i c_{2i}, \tag{8.41}$$

де перший доданок c_0 відображає деякі фіксовані витрати, які мають місце при проведенні обстеження, а інші три доданки – це змінні витрати. Зокрема, в цьому виразі c_u – вартість відбору однієї ПВО (кластера), c_{el} – вартість реєстрації одного елемента відібраного кластера, c_{2i} – вартість обстеження одного елемента i-го кластера, включаючи вартість інтерв'ю.

Функція витрат (8.41) цілком відображає загальні витрати на проведення обстеження з використанням двостадійного відбору елементів, але, на жаль, безпосередньо використати її для розв'язання нашої задачі оптимізації ми не можемо, оскільки ця функція є випадковою величиною – вона залежить від випадкової вибірки s_1. Тому використаємо математичне сподівання змінних витрат, тобто останніх трьох доданків виразу (8.41), яке позначимо через EVC (від англ. `expected variable cost`):

$$EVC := E(C(s_1)) = n_1 c_u + \frac{n_1}{N_1}\left(\sum_{i\in U_1} N_i\right) c_{el} + \frac{n_1}{N_1}\sum_{i\in U_1} n_i c_{2i},$$

або

$$EVC = a_1 f_1 + f_1\left(\sum_{i\in U_1} a_{2i} f_i\right), \tag{8.42}$$

де

$$a_1 = N_1 c_1; \quad a_{2i} = N_i c_{2i}; \quad c_1 = c_u + \overline{N} c_{el}. \tag{8.43}$$

Тут c_1 – середні витрати на один кластер, які включають витрати на відбір кластера та реєстрацію (створення списку) елементів у ньому. Коефіцієнт c_{2i} відображає витрати на один елемент i-го кластера, які включають витрати на відбір елемента та його обстеження.

Нагадаємо, в чому саме полягає задача оптимізації вибіркового дизайну: потрібно знайти такі значення f_1 та f_i, при яких

1) дисперсія (8.38) буде мінімальною при фіксованих загальних витратах (8.42):

$$EVC = C_0 \tag{8.44}$$

або

2) загальні витрати (8.42) будуть мінімальними при фіксованому значенні дисперсії:

$$\mathcal{D}_{\text{ПВДВЕ}}(\widehat{t}_\pi) = V_0. \tag{8.45}$$

Зосередимося на випадку, коли виконується умова $G \geq 0$ (тобто $A_1 \geq 0$)[6]. Тоді задача оптимізації вибіркового дизайну розв'язується досить легко, якщо використати нерівність Коші або метод невизначених множників Лагранжа[7]. Якщо тимчасово не брати до уваги обмеження (8.36) та (8.37), то за будь-яким із цих методів можна легко показати, що оптимальні значення часток відбору другої стадії як для випадку 1), так і для випадку 2) становлять:

$$f_i = \left(\frac{a_1}{A_1}\frac{A_{2i}}{a_{2i}}\right)^{1/2}, \quad i \in U_1. \tag{8.46}$$

Тепер, наприклад, для того, щоб знайти оптимальне значення частки відбору першої стадії для випадку 1), потрібно розв'язати рівняння (8.44) відносно f_1. Будемо мати

$$f_1 = \frac{C_0}{a_1}\left[1 + \sum_{i\in U_1}\left(\frac{a_{2i}A_{2i}}{a_1 A_1}\right)^{1/2}\right]^{-1}. \tag{8.47}$$

Аналогічно, для того, щоб знайти оптимальне значення частки відбору першої стадії для випадку 2), потрібно розв'язати рівняння (8.45) відносно f_1. Тоді

$$f_1 = \frac{A_1 + (A_1/a_1)^{1/2}\sum\limits_{i\in U_1}(a_{2i}A_{2i})^{1/2}}{V_0 - A_0}. \tag{8.48}$$

Якщо підставити у формули (8.46) та (8.47) вирази для обчислення коефіцієнтів A_1, A_{2i}, a_1 та a_{2i}, то отримаємо таке твердження.

[6] Від'ємне значення коефіцієнта A_1 теж можливе, хоча й малоймовірне.

[7] У загальному випадку розв'язати цю задачу оптимізації можна, використовуючи методи математичного програмування.

Твердження 8.3. *Нехай $G > 0$. Тоді дисперсія (8.38) буде мінімальною при фіксованих загальних витратах $EVC = C_0$, якщо:*

$$f_{\mathrm{I}} = \frac{n_{\mathrm{I}}}{N_{\mathrm{I}}} = \frac{C_0}{c_1}\left[N_{\mathrm{I}} + \frac{1}{(c_1 G)^{1/2}}\sum_{i\in U_{\mathrm{I}}} N_i S_i (c_{2i})^{1/2}\right]^{-1} \tag{8.49}$$

та

$$f_i = \frac{n_i}{N_i} = \left(\frac{c_1}{c_{2i}}\right)^{1/2}\frac{S_i}{G^{1/2}}, \quad i \in U_{\mathrm{I}}, \tag{8.50}$$

за умови, що виконуються обмеження (8.36) та (8.37).

Аналогічне твердження можна сформулювати і для випадку фіксованої дисперсії (вправа).

Зауваження 8.3. Якщо з виразу (8.49) випливає, що $n_{\mathrm{I}} \geq N_{\mathrm{I}}$, то це означає, що на першій стадії простого випадкового двостадійного відбору потрібно вибрати всі кластери генеральної сукупності і з кожного з них відбирати елементи згідно з формулою (8.50). Інакше кажучи, таке розміщення відповідає стратифікованому простому випадковому відбору.

8.4. Багатостадійний відбір

Незважаючи на складність багатостадійного відбору, він досить часто використовується при проведенні вибіркових обстежень великого масштабу. Як і у випадку двостадійного відбору, генеральну сукупність $U = \{1, 2, ..., k, ..., N\}$ ділять на N_{I} первинних вибіркових одиниць $U_1, U_2, ..., U_{N_{\mathrm{I}}}$. Позначимо множину первинних вибіркових одиниць через $U_{\mathrm{I}} = \{1, 2, ..., i, ..., N_{\mathrm{I}}\}$. Розмір N_i первинної вибіркової одиниці U_i відомий до початку обстеження, $i \in U_{\mathrm{I}}$.

Твердження 8.4. *У випадку r-стадійного відбору ($r \geq 2$) незміщена оцінка сумарного значення досліджуваної характеристики генеральної сукупності має вигляд $\widehat{t} = \sum\limits_{i\in s_{\mathrm{I}}} \widehat{t}_i/\pi_{\mathrm{I}i}$, де $E(\widehat{t}_i|s_{\mathrm{I}}) = T_i$.*

Дисперсія оцінки сумарного значення популяції дорівнює

$$\mathcal{D}(\widehat{t}) = \sum_{i\in U_{\mathrm{I}}}\sum_{j\in U_{\mathrm{I}}} (\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})\frac{T_i}{\pi_{\mathrm{I}i}}\frac{T_j}{\pi_{\mathrm{I}j}} + \sum_{i\in U_{\mathrm{I}}}\frac{\mathcal{D}_i}{\pi_{\mathrm{I}i}},$$

де перший доданок відображає дисперсію першої стадії відбору, а другий є комбінацією дисперсій, спричинених всіма наступними стадіями відбору.

Незміщена оцінка дисперсії $\mathcal{D}(\widehat{t})$ обчислюється за формулою

$$\widehat{\mathcal{D}}(\widehat{t}) = \sum_{i\in s_{\mathrm{I}}}\sum_{j\in s_{\mathrm{I}}} \frac{(\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j})}{\pi_{\mathrm{I}ij}}\frac{\widehat{t}_i}{\pi_{\mathrm{I}i}}\frac{\widehat{t}_j}{\pi_{\mathrm{I}j}} + \sum_{i\in s_{\mathrm{I}}}\frac{\widehat{\mathcal{D}}_i}{(\pi_{\mathrm{I}i})^2},$$

де $E(\widehat{\mathcal{D}}_i|s_{\mathrm{I}}) = \mathcal{D}_i$ для всіх $i \in s_{\mathrm{I}}$.

Доведення цього твердження аналогічне доведенню твердження 8.2.

8.5. Вправи та питання для самоконтролю

8.1. Чим поділ генеральної сукупності на страти відрізняється від поділу сукупності на кластери?

8.2. Наскільки ефективним є простий випадковий одностадійний кластерний відбір порівняно з простим випадковим відбором без повернення?

8.3. Які переваги використання кластерного відбору ви знаєте?

8.4. Використовуючи властивості π-оцінки Горвіца–Томпсона, довести твердження 8.1.

8.5. Проводиться вибіркове обстеження, метою якого є оцінювання сумарного доходу домогосподарств деякого району міста. Цей район складається з 60 кварталів різного розміру. Загальна кількість домогосподарств у ньому дорівнює 5000. За допомогою простого випадкового відбору без повернення вибрано три квартали, в кожному з яких обстежено всі домогосподарства. Результати обстеження наведено нижче в таблиці.

Номер кварталу	Кількість домогосподарств у кварталі	Сумарний дохід домогосподарств кварталу
1	120	2100
2	100	2000
3	80	1500

1) Оцінити сумарний дохід домогосподарств району, використовуючи оцінку Горвіца–Томпсона.

2) Обчислити значення незміщеної оцінки дисперсії оцінки Горвіца–Томсона сумарного доходу домогосподарств району.

8.6. Використовуючи твердження 8.2, вивести формули (8.33)–(8.35) для оцінювання сумарного значення досліджуваної характеристики популяції, дисперсії отриманої оцінки, а також для оцінювання цієї дисперсії у випадку простого випадкового двостадійного відбору елементів.

8.7. Проводиться вибіркове обстеження з використанням простого випадкового двостадійного відбору елементів, метою якого є оцінювання сумарного значення характеристики y деякої генеральної сукупності. На першій стадії отримано просту випадкову вибірку s_{I} розміру $n_{\text{I}} = 5$ з $N_{\text{I}} = 50$ первинних вибіркових одиниць (кластерів). Із кожного кластера, що потрапив до вибірки s_{I}, отримано просту випадкову вибірку s_i розміру $n_i = 3$ з N_i елементів, $i \in s_{\text{I}}$. Результати обстеження наведено в таблиці.

i	N_i	y_k
19	5	41, 49, 49
45	8	49, 49, 45
47	5	31, 31, 35
50	9	39, 41, 61
31	7	49, 51, 33

1) Обчислити π-оцінку сумарного значення характеристики y для генеральної сукупності.

2) Обчислити значення незміщеної оцінки дисперсії та коефіцієнта варіації π-оцінки сумарного значення характеристики y.

8.8. Вивести формули для обчислення оптимального розміщення у випадку простого випадкового двостадійного відбору за умови фіксованої дисперсії, використовуючи значення коефіцієнтів A_1, A_{2i}, a_1 та a_{2i} ($i \in U_1$).

8.9. Банк обслуговує 39800 клієнтів. Інформація про кожного клієнта міститься у базі даних банку в окремому файлі. Файли розміщені у 3980 папках по 10 файлів у кожній папці. Потрібно оцінити частку клієнтів, яким банк надав кредит. Для цього за допомогою простого випадкового відбору вибрано 40 папок (вибірка s). У кожній із вибраних папок підраховано кількість клієнтів (A_i), яким банк надав кредит, $i = \overline{1, 40}$. У результаті отримано такі дані:

$$\sum_{i \in s} A_i = 185, \quad \sum_{i \in s} A_i^2 = 1263.$$

1) Як називається такий метод відбору?

2) Записати вираз для точного обчислення частки клієнтів, яким банк надав кредит, та обчислити незміщену оцінку цього параметра.

3) Оцінити дисперсію отриманої оцінки і побудувати 95-відсотковий довірчий інтервал для оцінюваного параметра.

4) Обчислити дизайн-ефект для цього методу відбору.

Розділ 9

Оцінювання функцій від сумарних значень характеристик генеральної сукупності

9.1. Оцінювання вектора сумарних значень

У більшості вибіркових обстежень зазвичай досліджуються кілька характеристик генеральної сукупності. Розглянемо випадок, коли сумарні значення кількох характеристик (змінних) оцінюються за допомогою відповідних π-оцінок.

Припустимо, що досліджуються q характеристик генеральної сукупності, які ми будемо позначати через $y_1, \dots, y_j, \dots, y_q$. Значення цих характеристик для N елементів генеральної сукупності будемо позначати через $y_{j1}, \dots, y_{jk}, \dots, y_{jN}$, $j = 1, \dots, q$. Потрібно оцінити q компонент вектора невідомих сумарних значень цих характеристик:

$$\mathbf{T} = (T_1, \dots, T_j, \dots, T_q)',$$

де

$$T_j = \sum_{k \in U} y_{jk}.$$

З генеральної сукупності U відбирається ймовірнісна вибірка s згідно з вибірковим планом $p(s)$ з імовірностями включення π_k та π_{kl}. Для кожного $k \in s$ спостерігається вектор

$$\mathbf{y}_k = (y_{1k}, \dots, y_{jk}, \dots, y_{qk}).$$

Нехай сумарне значення кожної досліджуваної характеристики оцінюється за допомогою оцінки Горвіца–Томпсона (π-оцінки). Тоді вектор оцінок сумарних значень буде мати вигляд

$$\widehat{\mathbf{t}}_\pi = (\widehat{t}_{1\pi}, \dots, \widehat{t}_{j\pi}, \dots, \widehat{t}_{q\pi})',$$

де

$$\widehat{t}_{j\pi} = \sum_{k \in s} \frac{y_{jk}}{\pi_k}.$$

Очевидно, що

$$E(\widehat{\mathbf{t}}_\pi) = \mathbf{T},$$

тобто вектор $\widehat{\mathbf{t}}_\pi$ є незміщеною оцінкою вектора $\mathbf{T}$.

***Твердження* 9.1.** *Нехай* $\widehat{\mathbf{t}}_\pi = (\widehat{t}_{1\pi}, \dots, \widehat{t}_{j\pi}, \dots, \widehat{t}_{q\pi})'$ *– це вектор* π*-оцінок, що відповідає* q *змінним* $y_1, \dots, y_j, \dots, y_q$, *де* $\widehat{t}_{j\pi} = \sum\limits_{k \in s} \frac{y_{jk}}{\pi_k}$. *Тоді коваріаційна матриця*

$$\mathbf{V}(\widehat{\mathbf{t}}_\pi) = E\left[(\widehat{\mathbf{t}}_\pi - \mathbf{T})(\widehat{\mathbf{t}}_\pi - \mathbf{T})'\right] \tag{9.1}$$

є симетричною матрицею, в якій j*-й діагональний елемент є дисперсією оцінки* $\widehat{t}_{j\pi}$*:*

$$\mathcal{D}(\widehat{t}_{j\pi}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_{jk}}{\pi_k} \frac{y_{jl}}{\pi_l}, \quad j = 1, \dots, q, \tag{9.2}$$

а елементи поза діагоналлю цієї матриці задають коваріацію оцінок $\widehat{t}_{i\pi}$ *та* $\widehat{t}_{j\pi}$ $(i, j = 1, \dots, q,\ i \neq j)$*:*

$$\mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) := \operatorname{cov}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_{ik}}{\pi_k} \frac{y_{jl}}{\pi_l}. \tag{9.3}$$

Незміщеною оцінкою матриці $\mathbf{V}(\widehat{\mathbf{t}}_\pi)$ *є матриця* $\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_\pi)$, *діагональними елементами якої є оцінки дисперсій*

$$\widehat{\mathcal{D}}(\widehat{t}_{j\pi}) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_{jk}}{\pi_k} \frac{y_{jl}}{\pi_l}, \quad j = 1, \dots, q, \tag{9.4}$$

а поза діагоналлю стоять оцінки коваріацій

$$\widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_{ik}}{\pi_k} \frac{y_{jl}}{\pi_l}, \ i, j = \overline{1, q} : i \neq j. \tag{9.5}$$

Доведення. Формули (9.2) та (9.4) для обчислення діагональних елементів матриць $\mathcal{D}(\widehat{\mathbf{t}}_\pi)$ та $\widehat{\mathcal{D}}(\widehat{\mathbf{t}}_\pi)$ безпосередньо випливають із твердження 1.3.

Для $i, j = 1, \ldots, q : i \neq j$ будемо мати

$$\mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \operatorname{cov}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \operatorname{cov}\left(\sum_{k\in U} I_k \frac{y_{ik}}{\pi_k}, \sum_{k\in U} I_k \frac{y_{jk}}{\pi_k}\right) =$$

$$= \sum_{k\in U}\sum_{l\in U} \operatorname{cov}(I_k, I_l)\frac{y_{ik}}{\pi_k}\frac{y_{jl}}{\pi_l} = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{y_{ik}}{\pi_k}\frac{y_{jl}}{\pi_l},$$

тобто, маємо рівність (9.3).

Для того, щоб довести незміщеність оцінки (9.5) цієї коваріації, потрібно, як і у твердженні 1.3, скористатися тим фактом, що

$$E\left[I_k I_l\left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\right)\right] = \pi_{kl}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} = \pi_{kl} - \pi_k\pi_l,$$

звідки випливає, що

$$E[\widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi})] = E\left[\sum_{k\in s}\sum_{l\in s}\frac{(\pi_{kl} - \pi_k\pi_l)}{\pi_{kl}}\frac{y_{ik}}{\pi_k}\frac{y_{jl}}{\pi_l}\right] =$$

$$= E\left[\sum_{k\in U}\sum_{l\in U} I_k I_l\frac{(\pi_{kl} - \pi_k\pi_l)}{\pi_{kl}}\frac{y_{ik}}{\pi_k}\frac{y_{jl}}{\pi_l}\right] =$$

$$= \sum_{k\in U}\sum_{l\in U} E\left[I_k I_l\frac{(\pi_{kl} - \pi_k\pi_l)}{\pi_{kl}}\right]\frac{y_{ik}}{\pi_k}\frac{y_{jl}}{\pi_l} =$$

$$= \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{y_{ik}}{\pi_k}\frac{y_{jl}}{\pi_l} = \mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}),$$

що і треба було довести. □

Приклад 9.1. Одним із застосувань твердження 9.1 є випадок оцінювання кількох сумарних значень, які базуються на одній і тій самій досліджуваній змінній y. Наприклад, нехай потрібно оцінити такі три сумарні значення:

$$T_1 = \sum_{k\in U} y_k, \quad T_2 = \sum_{k\in U} y_k^2, \quad T_3 = \sum_{k\in U} y_k^3.$$

Розв'язок цієї задачі випливає з твердження 9.1, якщо покласти $q = 3$, $y_{1k} = y_k$, $y_{2k} = y_k^2$ та $y_{3k} = y_k^3$, $k \in U$. ◊

9.2. Оцінювання функцій від сумарних значень кількох змінних

Розглянемо задачу оцінювання параметра θ генеральної сукупності U, який можна подати у вигляді функції від q сумарних значень $T_1, \ldots, T_q$:

$$\theta = f(T_1, \ldots, T_q),$$

де $T_j = \sum_{k\in U} y_{jk}$, $j = 1, \ldots, q$.

Як і в попередньому параграфі, припустимо, що з генеральної сукупності U відбирається ймовірнісна вибірка s згідно з вибірковим планом $p(s)$ з імовірностями включення π_k та π_{kl}. Для кожного $k \in s$ спостерігається вектор $\mathbf{y}_k = (y_{1k}, \ldots, y_{jk}, \ldots, y_{qk})$, що дає можливість отримати π-оцінки сумарних значень $T_1, \ldots, T_q$:

$$\widehat{t}_{j\pi} = \sum_{k\in s}\frac{y_{jk}}{\pi_k}, \ j = 1, \ldots, q.$$

Ідея, яка лежить в основі методу оцінювання параметра θ, полягає в тому, щоб замість невідомих сумарних значень $T_1, \ldots, T_q$ підставити у функцію $f(\cdot, \ldots, \cdot)$ їх π-оцінки. Тобто оцінка параметра θ матиме вигляд

$$\hat{\theta} = f(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}). \tag{9.6}$$

9.2.1. Оцінювання лінійних функцій від сумарних значень кількох змінних

Якщо функція f – лінійна, тобто

$$\theta = a_0 + \sum_{j=1}^{q} a_j T_j,$$

то властивості оцінки $\hat{\theta}$ можна легко дослідити. У цьому випадку оцінка

$$\hat{\theta} = a_0 + \sum_{j=1}^{q} a_j \widehat{t}_{j\pi} \tag{9.7}$$

є незміщеною оцінкою параметра θ, а дисперсія цієї оцінки

$$\mathcal{D}(\hat{\theta}) = \mathcal{D}\left(\sum_{j=1}^{q} a_j \widehat{t}_{j\pi}\right) = \sum_{i=1}^{q}\sum_{j=1}^{q} a_i a_j \mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}), \qquad (9.8)$$

де коваріація $\mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi})$ визначається за формулою (9.3). Якщо $i = j$, то $\mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \mathcal{D}(\widehat{t}_{j\pi})$.

Оцінкою дисперсії (9.8) є статистика

$$\widehat{\mathcal{D}}(\hat{\theta}) = \sum_{i=1}^{q}\sum_{j=1}^{q} a_i a_j \widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}), \qquad (9.9)$$

де оцінка $\widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi})$ визначається за формулою (9.5). Якщо $i = j$, то $\widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \widehat{\mathcal{D}}(\widehat{t}_{j\pi})$.

9.2.2. Оцінювання нелінійних функцій від сумарних значень кількох змінних. Метод лінеаризації Тейлора

Тепер розглянемо випадок, коли $\theta = f(T_1, \ldots, T_q)$ – нелінійна функція від q сумарних значень $T_1, \ldots, T_q$. У цьому випадку зазвичай неможливо отримати точні значення зміщення та дисперсії оцінки $\hat{\theta} = f(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi})$. Для того, щоб спростити це завдання, використовується *метод лінеаризації Тейлора*, за допомогою якого можна вивести наближені формули для дисперсії оцінки $\hat{\theta}$ та оцінки цієї дисперсії. Також за допомогою цього методу можна будувати наближені довірчі інтервали для параметра θ.

Метод лінеаризації Тейлора полягає у наближенні нелінійної оцінки $\hat{\theta}$ псевдооцінкою $\hat{\theta}_0$, яка є лінійною функцією від змінних $\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}$. Якщо це наближення досить точне, то замість дисперсії $\mathcal{D}(\hat{\theta})$ можна використовувати дисперсію $\mathcal{D}(\hat{\theta}_0)$, яка обчислюється порівняно легко.

Для того, щоб знайти $\hat{\theta}_0$, використовується розклад функції f у ряд Тейлора в околі точки $(T_1, \ldots, T_q)$, в якому нехтують доданками другого і вище порядків. Отримаємо

$$\hat{\theta} \approx \hat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j (\widehat{t}_{j\pi} - T_j), \qquad (9.10)$$

де

$$a_j = \left.\frac{\partial f}{\partial \widehat{t}_{j\pi}}\right|_{(\widehat{t}_{1\pi},\ldots,\widehat{t}_{q\pi})=(T_1,\ldots,T_q)}. \qquad (9.11)$$

Для вибірок великого розміру оцінка $\hat{\theta}$ має приблизно такі самі властивості, як і лінійна статистика $\hat{\theta}_0$. Тому спробуємо наблизити зміщення та дисперсію оцінки $\hat{\theta}$ відповідними характеристиками оцінки $\hat{\theta}_0$.

Позначимо наближене значення дисперсії оцінки $\hat{\theta}$ через $\widetilde{\mathcal{D}}(\hat{\theta})$, і це наближене значення дорівнює точному значенню дисперсії лінеаризованої оцінки $\hat{\theta}_0$: $\widetilde{\mathcal{D}}(\hat{\theta}) = \mathcal{D}(\hat{\theta}_0)$.

Для зручності запису подальших формул введемо таке позначення:

$$u_k = \sum_{j=1}^{q} a_j y_{jk}. \qquad (9.12)$$

Тоді наближене значення дисперсії оцінки $\hat{\theta}$ обчислюється так:

$$\widetilde{\mathcal{D}}(\hat{\theta}) = \mathcal{D}(\hat{\theta}_0) = \mathcal{D}\left(\sum_{j=1}^{q} a_j \widehat{t}_{j\pi}\right) =$$

$$= \mathcal{D}\left(\sum_{k\in s} \frac{u_k}{\pi_k}\right) = \sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l)\frac{u_k}{\pi_k}\frac{u_l}{\pi_l}. \qquad (9.13)$$

Зауваження 9.1. Дисперсію $\mathcal{D}(\hat{\theta}_0)$ також можна розглядати як наближення середньоквадратичної похибки оцінки $\hat{\theta}$. Оскільки $E(\hat{\theta}_0) = \theta$, то

$$MSE(\hat{\theta}) \approx MSE(\hat{\theta}_0) = \mathcal{D}(\hat{\theta}_0).$$

Тепер проблема полягає в тому, щоб оцінити дисперсію (9.13). Величини u_k залежать від коефіцієнтів $a_1, \ldots, a_q$, які, в свою чергу, залежать від невідомих сумарних значень $T_1, \ldots, T_q$. Отже, u_k – невідомі. Стандартний вихід із цієї ситуації полягає в заміні невідомих сумарних значень характеристик генеральної сукупності π-оцінками цих сумарних значень. У результаті отримаємо оцінки

$\hat{a}_1, \ldots, \hat{a}_q$ коефіцієнтів $a_1, \ldots, a_q$, що дає нам можливість для всіх $k \in s$ обчислити значення величин

$$\hat{u}_k = \sum_{j=1}^{q} \hat{a}_j y_{jk}. \tag{9.14}$$

Тоді оцінка дисперсії (9.13) буде мати вигляд

$$\widehat{\mathcal{D}}(\hat{\theta}) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}. \tag{9.15}$$

Обґрунтування цієї процедури полягає в тому, що оцінка $\hat{u}_k$ як функція від π-оцінок сумарних значень є конзистентною оцінкою величини u_k. Оцінка $\widehat{\mathcal{D}}(\hat{\theta})$ як функція від конзистентних оцінок $\hat{u}_k$ при великих розмірах вибірок поводить себе так, ніби вона ґрунтується на істинних значеннях u_k. Таким чином, можна припустити, що $\widehat{\mathcal{D}}(\hat{\theta})$ є конзистентною оцінкою дисперсії $\mathcal{D}(\hat{\theta})$. Насправді запропонована оцінка є оцінкою наближеної дисперсії $\widetilde{\mathcal{D}}(\hat{\theta})$, але при великих розмірах вибірок $\widetilde{\mathcal{D}}(\hat{\theta})$ та $\mathcal{D}(\hat{\theta})$ будуть приблизно однаковими, тому формулу (9.15) можна використовувати для оцінювання дисперсії $\mathcal{D}(\hat{\theta})$. Правомірність використання цієї оцінки дисперсії було підтверджено моделюванням різних випадків оцінювання нелінійних параметрів генеральної сукупності [20].

Зауваження 9.2. У наступному твердженні ми використовуємо вираз «*приблизно незміщена оцінка*», який слід розуміти так: зміщення запропонованої оцінки є настільки малим порівняно з її дисперсією, що ним можна знехтувати.

***Твердження* 9.2.** *Приблизно незміщеною оцінкою параметра*

$$\theta = f(T_1, \ldots, T_q), \quad \text{де } T_1 = \sum_{k \in U} y_{1k}, \ldots, T_q = \sum_{k \in U} y_{qk},$$

є оцінка

$$\hat{\theta} = f(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}),$$

де $\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}$ *– π-оцінки сумарних значень* $T_1, \ldots, T_q$. *Застосувавши метод лінеаризації Тейлора (формули (9.10)–(9.11)), отримаємо наближене значення дисперсії оцінки* $\hat{\theta}$:

$$\widetilde{\mathcal{D}}(\hat{\theta}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{u_k}{\pi_k} \frac{u_l}{\pi_l},$$

де $u_k = \sum_{j=1}^{q} a_j y_{jk}$, *а коефіцієнти* a_j *визначені в (9.11).*

Оцінку цієї дисперсії можна обчислити так:

$$\widehat{\mathcal{D}}(\hat{\theta}) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l},$$

де $\hat{u}_k = \sum_{j=1}^{q} \hat{a}_j y_{jk}$, *а коефіцієнти* $\hat{a}_j$ *отримані шляхом заміни невідомих сумарних значень відповідними π-оцінками.*

У випадку вибіркового дизайну фіксованого розміру можна застосувати альтернативну оцінку:

$$\widehat{\mathcal{D}}(\hat{\theta}) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \left(\frac{\hat{u}_k}{\pi_k} - \frac{\hat{u}_l}{\pi_l} \right)^2. \tag{9.16}$$

Зауваження 9.3. Слід мати на увазі, що застосування методу лінеаризації Тейлора у випадках вибірок невеликого розміру призводить до недооцінювання дисперсії, причому величина цього зміщення залежить від того, наскільки складною є оцінка досліджуваного параметра. Наприклад, у випадку такої простої статистики, як зважене вибіркове середнє $\widetilde{\bar{y}} = \left(\sum_{k \in s} y_k / \pi_k \right) / \left(\sum_{k \in s} 1/\pi_k \right)$, недооцінюванням дисперсії можна знехтувати навіть у випадку невеликої вибірки. А от у випадках таких статистик, як оцінка дисперсії досліджуваної характеристики генеральної сукупності або коваріації двох характеристик сукупності, можуть знадобитися вибірки дуже великих розмірів, щоб можна було знехтувати цим зміщенням.

9.3. Оцінювання відношення сумарних значень двох досліджуваних характеристик

Розглянемо задачу оцінювання відношення (частки) невідомих сумарних значень двох досліджуваних характеристик генеральної сукупності:

$$R = \frac{T_y}{T_z} = \frac{\sum\limits_{k\in U} y_k}{\sum\limits_{k\in U} z_k}. \tag{9.17}$$

Наприклад, якщо U – це генеральна сукупність домогосподарств, y_k – сумарний дохід k-го домогосподарства, а z_k дорівнює кількості осіб в k-ому домогосподарстві, то R – це середній дохід у розрахунку на одну особу цієї генеральної сукупності.

Якщо невідомі сумарні значення оцінити за допомогою π-оцінок Горвіца–Томпсона: $\hat{t}_{y\pi} = \sum\limits_{k\in s} \frac{y_k}{\pi_k}$ та $\hat{t}_{z\pi} = \sum\limits_{k\in s} \frac{z_k}{\pi_k}$, то отримаємо оцінку відношення R, яка є нелінійною функцією від π-оцінок невідомих сумарних значень T_y та T_z:

$$\widehat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}. \tag{9.18}$$

Ця оцінка є зміщеною. Зміщення $B(\widehat{R}) = E(\widehat{R}) - R$ оцінки $\widehat{R}$ задовольняє таку умову (див. [20]):

$$\frac{[B(\widehat{R})]^2}{\mathcal{D}(\widehat{R})} = \frac{[E(\widehat{R}) - R]^2}{\mathcal{D}(\widehat{R})} \le \frac{\mathcal{D}(\hat{t}_{z\pi})}{T_z^2}. \tag{9.19}$$

Це означає, що якщо відносна стандартна похибка $\frac{\sqrt{\mathcal{D}(\hat{t}_{z\pi})}}{|T_z|}$ оцінки $\hat{t}_{z\pi}$ прямує до нуля зі збільшенням розміру вибірки (як зазвичай і буває), то відносне зміщення $\frac{B(\widehat{R})}{\sqrt{\mathcal{D}(\widehat{R})}}$ теж буде прямувати до нуля. Це дуже важлива властивість з точки зору побудови довірчих інтервалів.

Тепер застосуємо метод лінеаризації Тейлора, описаний у попередньому параграфі, для того, щоб знайти наближену дисперсію

оцінки $\widehat{R}$, а також оцінку цієї дисперсії. Оцінка $\widehat{R}$ є функцією від двох випадкових величин $\hat{t}_{y\pi}$ та $\hat{t}_{z\pi}$:

$$\widehat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}} = f(\hat{t}_{y\pi}, \hat{t}_{z\pi}).$$

Знайдемо частинні похідні, необхідні для оцінювання:

$$\frac{\partial \widehat{R}}{\partial \hat{t}_{y\pi}} = \frac{1}{\hat{t}_{z\pi}}, \quad \frac{\partial \widehat{R}}{\partial \hat{t}_{z\pi}} = -\frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}^2}.$$

Обчислимо значення цих частинних похідних в точці (T_y, T_z):

$$a_1 = \left.\frac{\partial \widehat{R}}{\partial \hat{t}_{y\pi}}\right|_{(T_y,T_z)} = \frac{1}{T_z};$$

$$a_2 = \left.\frac{\partial \widehat{R}}{\partial \hat{t}_{z\pi}}\right|_{(T_y,T_z)} = -\frac{T_y}{T_z^2} = -\frac{R}{T_z}.$$

Далі маємо

$$u_k = a_1 y_k + a_2 z_k = \frac{1}{T_z}(y_k - R z_k)$$

та

$$\hat{u}_k = \frac{1}{\hat{t}_{z\pi}}(y_k - \widehat{R} z_k).$$

Звідси випливає таке твердження.

***Твердження* 9.3.** *Результатом застосування методу лінеаризації Тейлора є таке лінійне наближення статистики* $\widehat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}$*:*

$$\widehat{R} \approx \widehat{R}_0 = R + \frac{1}{T_z} \sum_{k\in s} \frac{(y_k - R z_k)}{\pi_k}. \tag{9.20}$$

Оцінка $\widehat{R}$ *є приблизно незміщеною оцінкою параметра* R*, а наближене значення дисперсії цієї оцінки обчислюється за формулою*

$$\widetilde{\mathcal{D}}(\widehat{R}) = \mathcal{D}(\widehat{R}_0) = \tag{9.21}$$

$$= \frac{1}{T_z^2} \sum_{k\in U} \sum_{l\in U} (\pi_{kl} - \pi_k\pi_l) \frac{(y_k - Rz_k)}{\pi_k} \frac{(y_l - Rz_l)}{\pi_l}.$$

Оцінка дисперсії оцінки $\widehat{R}$ *має вигляд*

$$\widehat{\mathcal{D}}(\widehat{R}) = \frac{1}{\hat{t}_{z\pi}^2} \sum_{k\in s} \sum_{l\in s} \frac{(\pi_{kl} - \pi_k\pi_l)}{\pi_{kl}} \frac{(y_k - \widehat{R}z_k)}{\pi_k} \frac{(y_l - \widehat{R}z_l)}{\pi_l}. \qquad (9.22)$$

Зауваження 9.4. Інколи зручно користуватися такими виразами:

$$\widehat{R}_0 = R + \frac{1}{T_z}(\hat{t}_{y\pi} - R\hat{t}_{z\pi}),$$

$$\widetilde{\mathcal{D}}(\widehat{R}) = \frac{1}{T_z^2} \left(\mathcal{D}(\hat{t}_{y\pi}) + R^2\mathcal{D}(\hat{t}_{z\pi}) - 2R\mathbf{C}(\hat{t}_{y\pi}, \hat{t}_{z\pi})\right),$$

$$\widehat{\mathcal{D}}(\widehat{R}) = \frac{1}{\hat{t}_{z\pi}^2} \left(\widehat{\mathcal{D}}(\hat{t}_{y\pi}) + \widehat{R}^2\widehat{\mathcal{D}}(\hat{t}_{z\pi}) - 2\widehat{R}\widehat{\mathbf{C}}(\hat{t}_{y\pi}, \hat{t}_{z\pi})\right).$$

Зауваження 9.5. У випадку використання наближення (9.20) виконується рівність

$$E(\widehat{R}) \approx E(\widehat{R}_0) = R,$$

тобто, зміщенням оцінки $\widehat{R}$ ми нехтуємо. Але для вибірок малого розміру це наближення є досить грубим. Щоб отримати хоча б приблизну формулу для обчислення зміщення оцінки $\widehat{R}$, потрібно залишити у її розкладі в ряд Тейлора ще й члени другого порядку.

Приклад 9.2. Розглянемо простий випадковий відбір без повернення з розміром вибірки $n = fN$. Тоді $\hat{t}_{y\pi} = N\overline{y}$, $\hat{t}_{z\pi} = N\overline{z}$, $\widehat{R} = \frac{\overline{y}}{\overline{z}}$, де $\overline{y} = \frac{1}{n}\sum\limits_{k\in s} y_k$ та $\overline{z} = \frac{1}{n}\sum\limits_{k\in s} z_k$ – вибіркові середні характеристик y та z відповідно. Лінійне наближення (9.20) матиме вигляд

$$\widehat{R} \approx \widehat{R}_0 = R + \frac{1}{n\overline{Z}} \sum_{k\in s} (y_k - Rz_k) = R + \frac{\overline{y} - R\overline{z}}{\overline{Z}},$$

де $\overline{Z} = \frac{1}{N}\sum\limits_{k\in U} z_k$ – істинне середнє значення змінної z.

Вираз для наближеного обчислення дисперсії оцінки $\widehat{R}$ має вигляд

$$\widetilde{\mathcal{D}}(\widehat{R}) = \frac{1}{(\overline{Z})^2} \frac{1-f}{n} \frac{1}{N-1} \sum_{k\in U} (y_k - Rz_k)^2 =$$

$$= \frac{1}{(\overline{Z})^2} \frac{1-f}{n} \left(S_y^2 + R^2S_z^2 - 2RS_{yz}\right),$$

де S_{yz} – коваріація змінних y та z, обчислена за всією генеральною сукупністю.

Оцінка дисперсії оцінки $\widehat{R}$ дорівнює

$$\widehat{\mathcal{D}}(\widehat{R}) = \frac{1}{(\overline{z})^2} \frac{1-f}{n} \frac{1}{n-1} \sum_{k\in s} (y_k - \widehat{R}z_k)^2 =$$

$$= \frac{1}{(\overline{z})^2} \frac{1-f}{n} \left(\widehat{S}_y^2 + \widehat{R}^2\widehat{S}_z^2 - 2\widehat{R}\widehat{S}_{yz}\right),$$

де

$$\widehat{S}_y^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \overline{y})^2, \quad \widehat{S}_z^2 = \frac{1}{n-1}\sum_{k\in s}(z_k - \overline{z})^2,$$

$$\widehat{S}_{yz} = \frac{1}{n-1}\sum_{k\in s}(y_k - \overline{y})(z_k - \overline{z}).$$

◊

9.4. Оцінювання середнього значення характеристики генеральної сукупності

Розглянемо середнє значення досліджуваної характеристики y у розрахунку на один елемент генеральної сукупності. Цей параметр тісно пов'язаний із сумарним значенням $T_y = \sum\limits_{k\in U} y_k$ і визначається так:

$$\overline{Y} = \frac{1}{N}\sum_{k\in U} y_k = \frac{T_y}{N}. \qquad (9.23)$$

Якщо кількість елементів генеральної сукупності відома, тобто ми знаємо число N, то незміщеною оцінкою параметра $\overline{Y}$ є статистика

$$\widehat{\overline{y}}_\pi = \frac{\widehat{t}_{y\pi}}{N} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}. \tag{9.24}$$

Дисперсія цієї оцінки

$$\mathcal{D}(\widehat{\overline{y}}_\pi) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}, \tag{9.25}$$

а оцінка дисперсії має вигляд

$$\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \tag{9.26}$$

Альтернативний метод оцінювання середнього значення досліджуваної характеристики генеральної сукупності полягає у використанні оцінки сумарного значення T_y цієї характеристики та оцінки розміру генеральної сукупності N (незалежно від того, відомий він чи ні). Згідно з цим методом оцінка середнього значення обчислюється так:

$$\widetilde{\overline{y}} = \frac{\widehat{t}_{y\pi}}{\widehat{N}} = \frac{\sum\limits_{k\in s} y_k/\pi_k}{\sum\limits_{k\in s} 1/\pi_k}, \tag{9.27}$$

де $\widehat{N} = \sum\limits_{k\in s} 1/\pi_k$ – це π-оцінка параметра N.

Оцінка $\widetilde{\overline{y}}$ називається *зваженим вибірковим середнім*. Ця оцінка є нелінійною функцією від $\widehat{t}_{y\pi}$ та $\widehat{N}$, отже, вона не є незміщеною. Тому для дисперсії цієї оцінки ми можемо запропонувати тільки наближений вираз.

Якщо розглядати параметр $\overline{Y}$ як відношення двох сумарних значень: $\overline{Y} = \frac{T_y}{T_z}$, де $z = 1$ для всіх $k = 1, \ldots, N$, а оцінку $\widetilde{\overline{y}}$ – як відношення відповідних π-оцінок, то застосування твердження 9.3 до оцінки $\widehat{R} = \widetilde{\overline{y}}$ призводить до такого результату.

***Твердження* 9.4.** *Оцінка* $\widetilde{\overline{y}} = \frac{\sum\limits_{k\in s} y_k/\pi_k}{\sum\limits_{k\in s} 1/\pi_k}$ *є приблизно незміщеною оцінкою середнього значення досліджуваної характеристики генеральної сукупності. Наближене значення дисперсії цієї оцінки обчислюється за формулою*

$$\widehat{\mathcal{D}}(\widetilde{\overline{y}}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\left(\frac{y_k-\overline{Y}}{\pi_k}\right)\left(\frac{y_l-\overline{Y}}{\pi_l}\right). \tag{9.28}$$

Оцінка дисперсії оцінки $\widetilde{\overline{y}}$ *має вигляд*

$$\widehat{\mathcal{D}}(\widetilde{\overline{y}}) = \frac{1}{\widehat{N}^2}\sum_{k\in s}\sum_{l\in s}\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\left(\frac{y_k-\widetilde{\overline{y}}}{\pi_k}\right)\left(\frac{y_l-\widetilde{\overline{y}}}{\pi_l}\right). \tag{9.29}$$

Для деяких вибіркових дизайнів π-оцінка $\widehat{\overline{y}}_\pi$ дорівнює зваженому вибірковому середньому $\widetilde{\overline{y}}$. Наприклад, це виконується при ПВВбП та СТПВВ. Якщо число N невідоме, то тоді немає з чого вибирати, оскільки обчислити можна тільки оцінку $\widetilde{\overline{y}}$. Однак, якщо N відоме і ці дві оцінки відрізняються, то виникає потреба вибору однієї з них. Як це не дивно, навіть якщо N відоме заздалегідь, $\widetilde{\overline{y}}$ зазвичай є кращою оцінкою, ніж $\widehat{\overline{y}}_\pi$. Точні умови, за яких варто надати перевагу зваженому вибірковому середньому, сформулювати важко, тому наведемо лише кілька аргументів на користь використання цієї оцінки.

По-перше, з виразів для $\widehat{\mathcal{D}}(\widetilde{\overline{y}})$ та $\widehat{\mathcal{D}}(\widetilde{\overline{y}})$ видно, що оцінка $\widetilde{\overline{y}}$ тим краща, чим менші значення різниць $(y_k - \overline{Y})$, $k \in U$. По-друге, цю оцінку краще використовувати тоді, коли розмір вибірки змінний: якщо розмір вибірки більший, ніж очікувалося, то і чисельник, і знаменник у виразі для обчислення оцінки $\widetilde{\overline{y}}$ будуть мати більшу кількість доданків, а менший розмір вибірки відповідно призведе до меншої кількості доданків. Тобто, вплив розміру вибірки на відношення $\left(\sum\limits_{k\in s} y_k/\pi_k\right) / \left(\sum\limits_{k\in s} 1/\pi_k\right)$ буде певною мірою мінімальним.

Існує ще одна оцінка, яку варто використовувати у випадку

вибіркових дизайнів зі змінними розмірами вибірок:

$$\widehat{\overline{y}}_\pi^* = \frac{n}{Nn_s}\sum_{k\in s}\frac{y_k}{\pi_k} = \frac{n}{n_s}\widehat{\overline{y}}_\pi, \tag{9.30}$$

де $n = E(n_s) = \sum\limits_{k\in U}\pi_k$ – очікуваний розмір вибірки.

Якщо розмір вибірки фіксований, то $\widehat{\overline{y}}_\pi^* = \widehat{\overline{y}}_\pi$. Якщо всі ймовірності π_k однакові, $k = 1,\dots N$, то $\widehat{\overline{y}}_\pi^* = \widetilde{\overline{y}}$.

Оцінка $\widehat{\overline{y}}_\pi^*$ може мати значно меншу дисперсію, ніж $\widehat{\overline{y}}_\pi$ та $\widetilde{\overline{y}}$, якщо

1) розмір вибірки є випадковою величиною;

2) ймовірності π_k значно відрізняються;

3) відношення y_k/π_k приблизно однакове для всіх елементів генеральної сукупності.

9.5. Вправи та питання для самоконтролю

9.1. Які оцінки сумарних значень характеристик генеральної сукупності використовуються для оцінювання функцій від цих сумарних значень?

9.2. У чому полягає метод лінеаризації Тейлора? Яка мета його застосування?

9.3. Чи є оцінка відношення двох сумарних значень незміщеною?

9.4. Як оцінити середнє значення досліджуваної характеристики генеральної сукупності, якщо розмір сукупності невідомий?

9.5. Проводиться вибіркове обстеження, метою якого є оцінювання різниці між сумарними значеннями двох характеристик генеральної сукупності y та z. Розмір генеральної сукупності $N = 281$. За допомогою простого випадкового відбору без повернення отримано вибірку розміру $n = 40$, за якою обчислено такі величини:

$$\sum_{k\in s} y_k = 866;\ \widehat{S}_y^2 = 42,438;\ \sum_{k\in s} z_k = 383;\ \widehat{S}_z^2 = 29,430;\ \widehat{S}_{yz} = 1,309.$$

Побудувати 95-відсотковий довірчий інтервал для різниці сумарних значень змінних y та z.

9.6. Вивести альтернативні формули для обчислення дисперсії оцінки відношення у випадку вибіркового дизайну з фіксованим розміром вибірки, а також для оцінки цієї дисперсії.

9.7. З генеральної сукупності розміру $N = 124$ за допомогою відбору Бернуллі зі сталою ймовірністю включення $\pi = 0,3$ отримано вибірку розміру $n = 43$. Оцінити відношення сумарних значень двох характеристик генеральної сукупності y та z, якщо за вибірковими даними обчислено такі величини:

$$\sum_{k\in s} y_k = 669281;\quad \sum_{k\in s} z_k = 608902;$$

$$\sum_{k\in s}(y_k - \widehat{R}z_k)^2 = 3496228001,$$

де $\widehat{R} = \widehat{t}_{y\pi}/\widehat{t}_{z\pi}$. Знайти коефіцієнт варіації отриманої оцінки.

9.8. Довести, що у випадку використання відбору Бернуллі зважене вибіркове середнє дорівнює вибірковому середньому, тобто $\widetilde{\overline{y}} = \overline{y} = \sum\limits_{k\in s}\frac{y_k}{n_s}$.

9.9. Показати, що у випадку використання відбору Бернуллі з $\pi_k = \pi$ для всіх $k = 1,\dots,N$ наближене значення дисперсії зваженого вибіркового середнього $\widetilde{\overline{y}} = \overline{y}$ дорівнює $\widetilde{\mathcal{D}}(\overline{y}) = [(1-\pi)/N\pi]S_y^2$.

Розділ 10
Використання допоміжної інформації

10.1. Оцінювання за різницею

Із метою підвищення точності оцінок досліджуваних характеристик генеральної сукупності часто використовують наявну інформацію про інші змінні. *Допоміжна змінна* – це змінна, для якої наявна повна інформація до початку вибіркового обстеження. Інакше кажучи, значення допоміжної змінної заздалегідь відомі для кожного з N елементів генеральної сукупності.

В основі використання допоміжних змінних у процедурах оцінювання лежить припущення, що ці змінні корелюють із характеристикою, що вивчається. Розглянемо J допоміжних змінних $x_1, \ldots, x_j, \ldots, x_J$. Значення j-ї змінної для k-го елемента генеральної сукупності будемо позначати як x_{jk}. Для k-го елемента генеральної сукупності визначимо вектор значень допоміжних змінних: $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$.

Як і в попередніх розділах, позначимо досліджувану характеристику генеральної сукупності через y, а її значення для k-го елемента генеральної сукупності через y_k, $k \in U$. До початку вибіркового обстеження значення $y_1, \ldots, y_N$ залишаються невідомими, тоді як інформація про вектори $\mathbf{x}_1, \ldots, \mathbf{x}_N$ є у повному розпорядженні дослідника. Невідомий параметр генеральної сукупності, який потрібно оцінити, – це сумарне значення характеристики y:

$$T_y = \sum_{k \in U} y_k.$$

Вибірка s вибирається з U згідно з вибірковим дизайном $p(\cdot)$ з імовірностями включення $\pi_k > 0$ і $\pi_{kl} > 0$. Для кожного $k \in s$ ми маємо значення y_k і вектор значень допоміжних змінних $\mathbf{x}_k$. Завдання полягає в тому, щоб оцінити параметр T_y, маючи значення $(y_k, \mathbf{x}_k)$ для всіх $k \in s$, а також значення $\mathbf{x}_k$ для всіх $k \in U \setminus s$.

Головна ідея, яка лежить в основі оцінювання за різницею, полягає у використанні допоміжної інформації для створення N

наближених значень $y_1^0, \ldots, y_N^0$ характеристики y, таких що y_k^0 достатньо точно наближають значення y_k. Будемо формувати наближене значення y_k^0 у вигляді лінійної комбінації відомих значень допоміжних змінних $x_{1k}, \ldots, x_{Jk}$:

$$y_k^0 = \sum_{j=1}^{J} A_j x_{jk} = \mathbf{A}' x_k, \tag{10.1}$$

де $\mathbf{A} = (A_1, \ldots, A_J)'$ – вектор коефіцієнтів. Спочатку будемо припускати, що значення елементів цього вектора відомі. Очевидно, що тоді y_k^0 можна обчислити для всіх $k \in U$, бо значення допоміжних змінних відомі для всієї генеральної сукупності. Оскільки ми вважаємо, що y_k^0 достатньо точно наближає значення y_k, то це означає, що має місце і таке наближення:

$$y_k \approx \sum_{j=1}^{J} A_j x_{jk} = \mathbf{A}' x_k \,. \tag{10.2}$$

Використовуючи наближені значення характеристики y, невідоме сумарне значення цієї характеристики можна записати у такому вигляді:

$$T_y = \sum_{k \in U} y_k = \sum_{k \in U} y_k^0 + \sum_{k \in U} (y_k - y_k^0) = \sum_{k \in U} y_k^0 + \sum_{k \in U} D_k, \tag{10.3}$$

де $D_k = y_k - y_k^0$ для $k = 1, \ldots, N$. Наближене сумарне значення $\sum\limits_{k \in U} y_k^0$ у виразі (10.3) є відомою величиною, оскільки значення $y_1^0, \ldots, y_N^0$ – відомі. Але сума різниць $\sum\limits_{k \in U} D_k$ є невідомою, оскільки значення $y_1, \ldots, y_N$ – невідомі. Тому природно замінити у формулі (10.3) суму різниць відповідною сумою π-оцінок різниць:

$$\widehat{t}_{yD} = \sum_{k \in U} y_k^0 + \sum_{k \in s} \frac{D_k}{\pi_k}. \tag{10.4}$$

Оцінка $\widehat{t}_{yD}$ називається *оцінкою за різницею* (англ. `difference estimator`) сумарного значення генеральної сукупності.

***Твердження* 10.1.** *Оцінка за різницею (10.4) є незміщеною оцінкою сумарного значення* $T_y = \sum\limits_{k\in U} y_k$. *Дисперсія цієї оцінки обчислюється за формулою*

$$\mathcal{D}(\widehat{t}_{yD}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{D_k}{\pi_k}\frac{D_l}{\pi_l}. \qquad (10.5)$$

Незміщена оцінка дисперсії (10.5) обчислюється за формулою

$$\widehat{\mathcal{D}}(\widehat{t}_{yD}) = \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}}\frac{D_k}{\pi_k}\frac{D_l}{\pi_l}, \qquad (10.6)$$

де $\Delta_{kl} = (\pi_{kl} - \pi_k\pi_l)$, $k,l \in U : k \neq l$, *та* $\Delta_{kk} = \pi_k(1-\pi_k)$.

Доведення. Незміщеність оцінки за різницею (10.4) є наслідком того, що $\sum\limits_{k\in s}\frac{D_k}{\pi_k}$ є незміщеною оцінкою для $\sum\limits_{k\in U} D_k$.

Якщо зауважити, що

$$\mathcal{D}(\widehat{t}_{yD}) = \mathcal{D}\left(\sum_{k\in s}\frac{D_k}{\pi_k}\right),$$

то і формули (10.5) та (10.6) безпосередньо випливають із властивостей π-оцінки, наведених у розділі 1. □

Зауваження 10.1. Якщо вибірковий дизайн $p(s)$ має фіксований розмір n, то дисперсію (10.5) можна обчислювати і за альтернативною формулою

$$\mathcal{D}(\widehat{t}_{yD}) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\left(\frac{D_k}{\pi_k} - \frac{D_l}{\pi_l}\right)^2, \qquad (10.7)$$

а незміщена оцінка дисперсії $\mathcal{D}(\widehat{t}_{yD})$ може бути обчислена так:

$$\widehat{\mathcal{D}}(\widehat{t}_{yD}) = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{D_k}{\pi_k} - \frac{D_l}{\pi_l}\right)^2. \qquad (10.8)$$

Приклад 10.1. Якщо $y_k^0 = x_k$, то $D_k = y_k - x_k$. Тоді у випадку ПВВбП з розміром вибірки $n = fN$ дисперсія оцінки за різницею сумарного значення характеристики y буде мати вигляд

$$\mathcal{D}(\widehat{t}_{yD}) = N^2\frac{1-f}{n}\left(S_y^2 + S_x^2 - 2S_{xy}\right), \qquad (10.9)$$

де S_x^2 та S_y^2 – це дисперсії змінних x та y відповідно, обчислені за генеральною сукупністю; $S_{xy} = \frac{1}{N-1}\sum\limits_{k\in U}(x_k - \overline{X})(y_k - \overline{Y})$ – коваріація цих змінних; $\overline{X} = \frac{1}{N}\sum\limits_{k\in U} x_k$ та $\overline{Y} = \frac{1}{N}\sum\limits_{k\in U} y_k$ – середні значення, обчислені для всієї генеральної сукупності. Якщо кореляція $r = \frac{S_{xy}}{S_x S_y}$ велика, то дисперсія оцінки за різницею зазвичай значно менша, ніж дисперсія π-оцінки $\widehat{t}_{y\pi}$. Відношення дисперсій цих оцінок

$$\frac{\mathcal{D}(\widehat{t}_{yD})}{\mathcal{D}(\widehat{t}_{y\pi})} = 1 + \left(\frac{S_x}{S_y}\right)^2 - 2r\frac{S_x}{S_y}. \qquad (10.10)$$

У випадку ПВВбП оцінка за різницею з використанням $y_k^0 = x_k$ має меншу дисперсію, ніж дисперсія π-оцінки $\widehat{t}_{y\pi}$ тоді і тільки тоді, коли $r > \frac{S_x}{2S_y}$. ◊

Інший погляд на оцінку за різницею – це використання її як уточнення звичайної π -оцінки сумарного значення досліджуваної характеристики популяції:

$$\widehat{t}_{yD} = \widehat{t}_{y\pi} + \sum_{j=1}^{J} A_j(\widehat{t}_{x_j} - \widehat{t}_{x_j\pi}), \qquad (10.11)$$

де $\widehat{t}_{x_j\pi} = \sum\limits_{k\in s} x_{jk}/\pi_k$ – це π-оцінка сумарного значення $T_{x_j} = \sum\limits_{k\in U} x_{jk}$ допоміжної змінної x_j, $j = 1,\ldots,J$. Таким чином, оцінка за різницею (10.11) дорівнює сумі π-оцінки та коригуючого доданка.

10.2. Оцінювання за регресією

Розглянемо *метод оцінювання за регресією* з використанням J допоміжних змінних. Як і в попередньому параграфі, вектор значень допоміжної змінної для k-го елемента позначимо через $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$, $k = 1, \ldots, N$. Використовуючи значення $(y_k, \mathbf{x}_k)$ для всіх $k \in s$ та $\mathbf{x}_k$ для всіх $k \in U \setminus s$, потрібно оцінити сумарне значення $T_y = \sum\limits_{k \in U} y_k$.

Розпочнемо з оцінки за різницею виду (10.11):

$$\widehat{t}_{yD} = \widehat{t}_{y\pi} + \sum_{j=1}^{J} A_j (\widehat{t}_{x_j} - \widehat{t}_{x_j\pi}).$$

Якщо коефіцієнти $A_1, \ldots, A_J$ невідомі, то можна спробувати застосувати оцінку такого самого виду, але замість невідомих коефіцієнтів підставити їх оцінки. Оскільки оцінювання за різницею базується на припущенні про приблизну лінійну залежність між досліджуваною змінною y та допоміжними змінними $x_1, \ldots, x_J$ (див. (10.2)), то для оцінювання коефіцієнтів $A_1, \ldots, A_J$ можна застосувати *регресійний аналіз* [4].

Зауваження 10.2. Регресійний аналіз об'єднує велику кількість методів виявлення та дослідження статистичних залежностей між різними характеристиками об'єктів. Для ілюстрації базових понять регресійного аналізу розглянемо приклад з книги [4].

Проводиться серія з N дослідів над однотипними об'єктами, для кожного з яких вимірюються дві характеристики: X та Y. Результатом k-го досліду є пара (X_k, Y_k) значень цих характеристик для k-го об'єкта, а результатом серії – набір даних $(X_k, Y_k), k = 1, \ldots, N$. З певних теоретичних міркувань відомо, що між X та Y повинен бути функціональний зв'язок виду

$$Y = g(X; \beta), \tag{10.12}$$

де g – деяка відома функція, $\beta = (\beta_1, \ldots, \beta_m)'$ – вектор невідомих коефіцієнтів (параметрів), які є сталими у даній серії дослідів, хоча, взагалі кажучи, можуть набувати різних, часто апріорі невідомих значень. Однак, при підстановці реально виміряних характеристик (X_k, Y_k) у співвідношення (10.12) виявляється, що при жодному можливому значенні β ця рівність не виконується одночасно для всіх досліджуваних об'єктів.

Такі речі практики зазвичай пояснюють тим, що спостережувані дані завжди вимірюються із деякою похибкою. Часто буває, що значення X відоме досить точно, а от похибкою у вимірюванні Y знехтувати не можна. У цьому випадку справжнє значення Y у k-ому досліді має бути $g(X_k; \beta)$. Отже, $(Y_k - g(X_k; \beta))$ – це похибка у вимірюванні Y_k, яку ми позначимо через ε_k. У результаті маємо

$$Y_k = g(X_k; \beta) + \varepsilon_k. \tag{10.13}$$

Якщо коефіцієнти β відомі досліднику, то не виникає труднощів у знаходженні похибок ε_k. Складніші проблеми з'являються тоді, коли коефіцієнти насправді невідомі. Як оцінити їх, маючи неточні спостереження (X_k, Y_k)? Як перевіряти гіпотези, пов'язані з цими коефіцієнтами? Чи можна передбачити, які значення Y відповідатимуть заданим значенням X? Ці та деякі інші питання для даних, що описуються теоретичною моделлю (10.13), розв'язуються методами регресійного аналізу. Рівняння (10.13) називають *регресійним рівнянням*, параметри $\beta_1, \ldots, \beta_m$ – *параметрами* або *коефіцієнтами регресії*, а функцію g – *функцією регресії*.

У класичному регресійному аналізі розглядають лише моделі, в яких характеристика Y є дійсним числом, а X може бути як одним числом, так і вектором, тобто набором скалярних характеристик: $\mathbf{X} = (X_1, \ldots, X_J)$. Величину Y називають *залежною змінною* або *відгуком*, величини $X_1, \ldots, X_J$ – *незалежними змінними* або *регресорами*, а ε_k – *похибками* або *помилками регресії*.

Якщо X є скалярною характеристикою досліджуваних об'єктів, то пари (X_k, Y_k) можна трактувати як точки на площині, а рівняння (10.12) при різних значеннях параметра β задає сім'ю кривих на цій площині. Зображення експериментальних даних у вигляді точок на площині називають *діаграмою розсіювання*, а сім'ю кривих (10.12) – *теоретичними* або *регресійними кривими*.

Графічно задачу регресійного аналізу можна трактувати як підбір такої теоретичної кривої, яка проходить найближче до експериментальних точок.

Повернемося до задачі оцінювання за регресією.

Використаємо позначення, більш звичні для регресійного аналізу, і позначимо через $\widehat{B}_1, \ldots, \widehat{B}_J$ відповідні оцінки невідомих коефіцієнтів $A_1, \ldots, A_J$. Вибір оцінок $\widehat{B}_1, \ldots, \widehat{B}_J$ базується на припущенні про вигляд графічного зображення N точок скінченної генеральної сукупності U:

$$\{(y_k, x_{1k}, \ldots, x_{Jk}) : \ k = 1, \ldots, N\}, \tag{10.14}$$

а саме: припускається, що діаграма розсіювання (10.14) виглядає так, ніби вона отримана в результаті використання моделі лінійної регресії ξ, в якій досліджувана змінна y лінійно залежить від допоміжних змінних $x_1, \ldots, x_J$.

Регресійна модель ξ задовольняє такі умови:

1) $y_1, \ldots, y_N$ є реалізаціями незалежних випадкових величин $Y_1, \ldots, Y_N$;

2) $E_\xi(Y_k) = \sum\limits_{j=1}^{J} \beta_j x_{jk}$, $k = 1, \ldots, N$;

3) $\mathcal{D}_\xi(Y_k) = \sigma_k^2$, $k = 1, \ldots, N$,

де E_ξ і $\mathcal{D}_\xi$ – математичне сподівання та дисперсія відносно моделі ξ; $\beta_1, \ldots, \beta_J$ та $\sigma_1^2, \ldots, \sigma_N^2$ – параметри моделі ξ. Це типова модель регресійного аналізу.

Надалі не будемо відрізняти позначення випадкової величини Y від позначення її реалізації y.

Приклад 10.2. Наведемо два приклади моделі регресії з однією допоміжною змінною x:

$$\begin{cases} E_\xi(y_k) = \beta x_k, \\ \mathcal{D}_\xi(y_k) = \sigma^2 x_k, \end{cases} \tag{10.15}$$

(тут припускається, що значення $x_1, \ldots, x_N$ – додатні)

та

$$\begin{cases} E_\xi(y_k) = \beta_1 + \beta_2 x_k, \\ \mathcal{D}_\xi(y_k) = \sigma^2. \end{cases} \tag{10.16}$$

В обох цих моделях вважається, що $y_1, \ldots, y_N$ – незалежні випадкові величини. ◊

Означення 10.1. *Оцінка за регресією* (англ. `regression estimator`) визначається так [20]:

$$\widehat{t}_{yr} = \widehat{t}_{y\pi} + \sum_{j=1}^{J} \widehat{B}_j (T_{x_j} - \widehat{t}_{x_j\pi}), \tag{10.17}$$

де $\widehat{t}_{y\pi} = \sum\limits_{k\in s} y_k/\pi_k$ – π-оцінка сумарного значення характеристики y; $\widehat{t}_{x_j\pi} = \sum\limits_{k\in s} x_{jk}/\pi_k$ – π-оцінка відомого сумарного значення $T_{x_j} = \sum\limits_{k\in U} x_{jk}$ змінної x_j; $\widehat{B}_1, \ldots, \widehat{B}_J$ – компоненти вектора $\widehat{\mathbf{B}}$, який визначається за формулою

$$\widehat{\mathbf{B}} = (\widehat{B}_1, \ldots, \widehat{B}_J)' = \left(\sum_{k\in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k} \right)^{-1} \sum_{k\in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k}. \tag{10.18}$$

Оцінка $\widehat{t}_{yr}$ є приблизно незміщеною, і чим більший розмір вибірки, тим точнішою буде ця оцінка.

Зауваження 10.3. Як і у випадку оцінювання за різницею, оцінювання за регресією дає можливість покращити основну оцінку $\widehat{t}_{y\pi}$. Оцінка за регресією є сумою π-оцінки $\widehat{t}_{y\pi}$ та деякого коригуючого доданка. Якщо використання оцінки за регресією обґрунтоване, то коригуючий доданок зазвичай від'ємно корельований з похибкою π-оцінки. Якщо вибірка великого розміру та π-оцінка має велику похибку, але існує сильна лінійна залежність між досліджуваною та допоміжними змінними, то цей коригуючий доданок приблизно дорівнює похибці π-оцінки з протилежним знаком, тобто похибка оцінки $\widehat{t}_{yr}$ буде меншою, ніж похибка оцінки $\widehat{t}_{y\pi}$.

Зауваження 10.4. Роль моделі ξ полягає в тому, щоб «описати» графік розсіювання точок генеральної сукупності. Ми сподіваємося, що модель підібрана вдало, тобто, що генеральна сукупність виглядає так, ніби вона згенерована за допомогою моделі ξ. Але це зовсім не означає, що генеральна сукупність справді згенерована за допомогою цієї моделі, тому й параметри генеральної сукупності не залежать від вибору моделі. Модель ξ необхідна для того, щоб отримати прийнятні оцінки $\widehat{\mathbf{B}}$, потрібні для обчислення оцінки за регресією. Ефективність оцінки за регресією (порівняно з π-оцінкою) залежить від того, наскільки добре підібрана модель, але її основні властивості (наприклад, приблизна незміщеність) від цього не залежать. Це приклад *«оцінювання за допомогою моделі»* на відміну від *«оцінювання на основі моделі»*, при якому від моделі залежать і властивості оцінки.

Наведемо міркування, за допомогою яких отримано формулу (10.18) для обчислення оцінок $\widehat{B}_1, \dots, \widehat{B}_J$. Для цього розглянемо гіпотетичне суцільне обстеження генеральної сукупності, в якому для всіх $k \in U$ спостерігаються значення y_k та вектори $\mathbf{x}_k$. У цьому випадку оцінка за методом найменших квадратів вектора $\beta = (\beta_1, \dots, \beta_J)'$ параметрів моделі ξ матиме вигляд

$$\mathbf{B} = (B_1, \dots, B_J)' = \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2} \right)^{-1} \sum_{k \in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2}. \qquad (10.19)$$

Це відомий результат з регресійного аналізу, де також доведено, що вектор $\mathbf{B}$ є найкращою лінійною незміщеною оцінкою вектора β для даної моделі регресії.

Для нас вектор $\mathbf{B}$ залишається невідомим параметром скінченної генеральної сукупності U. Щоб оцінити його за вибіркою s, застосуємо підхід із використанням π-оцінок сумарних значень, описаний у розділі 9. Для цього запишемо вектор $\mathbf{B}$ в такому вигляді:

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{t}, \qquad (10.20)$$

де

$$\mathbf{T} = \sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2}, \quad \mathbf{t} = \sum_{k \in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2}.$$

Тут $\mathbf{T}$ – симетрична матриця розміру $J \times J$ з елементами

$$t_{ij} = \sum_{k \in U} \frac{x_{ik} x_{jk}}{\sigma_k^2} = t_{ji},$$

а $\mathbf{t}$ – J-вимірний вектор з компонентами

$$t_j = \sum_{k \in U} \frac{x_{jk} y_k}{\sigma_k^2}.$$

Очевидно, що незміщеною оцінкою матриці $\mathbf{T}$ є матриця

$$\widehat{\mathbf{T}} = \sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k},$$

елементами якої є π-оцінки елементів матриці $\mathbf{T}$, які обчислюються так:

$$\widehat{t}_{ij,\pi} = \sum_{k \in s} \frac{x_{ik} x_{jk}}{\sigma_k^2 \pi_k}, \quad i, j = \overline{1, J}.$$

Аналогічно, незміщеною оцінкою вектора $\mathbf{t}$ є вектор $\widehat{\mathbf{t}} = \sum_{k \in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k}$, компонентами якого є π-оцінки компонентів вектора $\mathbf{t}$, які обчислюються за формулою:

$$\widehat{t}_{j,\pi} = \sum_{k \in s} \frac{x_{jk} y_k}{\sigma_k^2 \pi_k}, \quad j = \overline{1, J}.$$

Згідно з методом використання π-оцінок для оцінювання функцій від сумарних значень характеристик генеральної сукупності, наведеним у розділі 9, параметр $\mathbf{B}$ генеральної сукупності тепер можна оцінити за допомогою статистики:

$$\widehat{\mathbf{B}} = (\widehat{B}_1, \dots, \widehat{B}_J)' = \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}} = \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k} \right)^{-1} \sum_{k \in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k},$$

тобто, ми вивели формулу (10.18).

Таким чином, ми показали, що $\widehat{\mathbf{B}}$ є оцінкою невідомого параметра $\mathbf{B}$, який, у свою чергу, є оцінкою параметра β регресійної моделі ξ за результатами гіпотетичного суцільного обстеження сукупності.

Зауваження 10.5. 1) Оцінка $\widehat{\mathbf{B}}$ є зміщеною оцінкою параметра $\mathbf{B}$. При великих розмірах вибірок зміщенням можна знехтувати.
2) Припускається, що матриці $\mathbf{T}$ та $\widehat{\mathbf{T}}$ невироджені, тому існують матриці, обернені до них.
3) Для того, щоб на основі вибіркових даних можна було обчислити оцінку $\widehat{\mathbf{B}}$ за формулою (10.18), потрібно, щоб вона залежала тільки від відомих параметрів. Тому параметри $\sigma_1^2, \ldots, \sigma_N^2$ мають задовольняти певні умови. Наприклад, можна вимагати, щоб всі σ_k^2 були відомими або щоб $\sigma_k^2 = \nu_k \sigma^2$ з невідомим параметром σ^2 та відомими коефіцієнтами $\nu_1, \ldots, \nu_N$ (спеціальний випадок: $\nu_k = 1$ для всіх $k \in U$). Існують й інші важливі випадки, в яких дисперсія залежить від кількох невідомих параметрів, які у виразі для обчислення оцінки $\widehat{\mathbf{B}}$ скорочуються.
4) Оскільки значення допоміжних змінних вважаються відомими для всіх елементів генеральної сукупності, то при відомих значеннях σ_k^2 можна знайти точні значення елементів матриці $\mathbf{T}$. Однак для обчислення оцінки $\widehat{\mathbf{B}}$ краще використовувати оцінку $\widehat{\mathbf{T}}$ цієї матриці.

Альтернативні формули для обчислення оцінки $\widehat{t}_{yr}$ мають такий вигляд:

$$\widehat{t}_{yr} = \sum_{k \in U} \widehat{y}_k + \sum_{k \in s} e_{ks}/\pi_k \tag{10.21}$$

та

$$\widehat{t}_{yr} = \sum_{k \in U} y_k^0 + \sum_{k \in s} g_{ks} E_k/\pi_k, \tag{10.22}$$

де $\widehat{y}_k = \mathbf{x}_k' \widehat{\mathbf{B}}$, $k \in U$, – наближені значення характеристики y, отримані в результаті використання *вибіркових даних* для наближення регресійної моделі ξ, а $e_{ks} = y_k - \widehat{y}_k$, $k \in s$, – відповідні залишки регресії; $y_k^0 = \mathbf{x}_k' \mathbf{B}$, $k \in U$, – наближені значення характеристики y, отримані в результаті використання *гіпотетичної*

генеральної сукупності для наближення регресійної моделі ξ, а $E_k = y_k - y_k^0$, $k \in s$, – відповідні залишки регресії у цьому випадку. Вагові коефіцієнти g_{ks} обчислюються за формулою

$$g_{ks} = 1 + \left(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} \mathbf{x}_k/\pi_k \right)' \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k'/\sigma_k^2 \pi_k \right)^{-1} \frac{\mathbf{x}_k}{\sigma_k^2}. \tag{10.23}$$

За допомогою розкладу в ряд Тейлора оцінку за регресією $\widehat{t}_{yr}$ (виду (10.17) чи (10.21)) можна наблизити оцінкою

$$\widetilde{t}_{yr} = \sum_{k \in U} y_k^0 + \sum_{k \in s} E_k/\pi_k,$$

звідки випливає формула для знаходження наближеного значення дисперсії оцінки $\widehat{t}_{yr}$:

$$\widetilde{\mathcal{D}}(\widehat{t}_{yr}) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{E_k E_l}{\pi_k \pi_l}. \tag{10.24}$$

Якщо в цій формулі замість значень E_k/π_k, яких ми насправді не знаємо, підставити значення e_{ks}/π_k, які можна обчислити за вибіркою, то отримаємо одну з формул для оцінювання дисперсії оцінки за регресією:

$$\widehat{\mathcal{D}}(\widehat{t}_{yr}) = \sum_{k \in U} \sum_{l \in U} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_{ks} e_{ls}}{\pi_k \pi_l}. \tag{10.25}$$

Із формули (10.22) маємо $\mathcal{D}(\widehat{t}_{yr}) = \mathcal{D}\left(\sum_{k \in s} \frac{g_{ks} E_k}{\pi_k} \right)$, де вираз у дужках нагадує π-оцінку сумарного значення деякої характеристики (див. (1.4)). Звідси випливає ще одна формула для оцінювання дисперсії оцінки $\widehat{t}_{yr}$:

$$\widehat{\mathcal{D}}(\widehat{t}_{yr}) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{g_{ks} e_{ks}}{\pi_k} \right) \left(\frac{g_{ls} e_{ls}}{\pi_l} \right). \tag{10.26}$$

Оцінки (10.25) та (10.26) будуть приблизно незміщеними при великих розмірах вибірок.

10.3. Оцінювання за відношенням

Нехай для оцінювання сумарного значення досліджуваної характеристики y генеральної сукупності використовується інформація про одну допоміжну змінну x, яка набуває тільки додатних значень. Значення $x_1, \dots, x_k, \dots x_N$ допоміжної змінної x для елементів генеральної сукупності відомі заздалегідь.

Модель регресії, яка базується на припущенні, що відношення y_k/x_k є приблизно константою, називається *моделлю відношення*. У цій моделі стверджується, що

$$E_\xi(y_k) = \beta x_k,$$

тобто, в даному випадку лінія регресії – пряма, яка проходить через початок координат. Для того, щоб ця модель була повністю визначена, потрібно ще зробити припущення щодо структури дисперсії $\mathcal{D}_\xi(y_k)$. Нехай ця дисперсія зростає пропорційно до змінної x. Тепер модель відношення визначена повністю і має вигляд

$$\begin{cases} E_\xi(y_k) = \beta x_k, \\ \mathcal{D}_\xi(y_k) = \sigma^2 x_k, \end{cases} \tag{10.27}$$

де значення параметрів β та σ^2 невідомі.

Оцінка за регресією, яка ґрунтується на моделі (10.27), називається *оцінкою за відношенням* (англ. `ratio estimator`). У наступному твердженні наведено властивості цієї оцінки.

***Твердження* 10.2.** *Оцінка за відношенням сумарного значення T_y характеристики y генеральної сукупності має такий вигляд:*

$$\widehat{t}_{yR} = \left(\sum_{k\in U} x_k\right) \frac{\sum\limits_{k\in s} y_k/\pi_k}{\sum\limits_{k\in s} x_k/\pi_k} = T_x \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}} = T_x\, \widehat{B}. \tag{10.28}$$

Ця оцінка є приблизно незміщеною оцінкою параметра T_y.

Наближене значення дисперсії оцінки $\widehat{t}_{yR}$ обчислюється за формулою

$$\widetilde{\mathcal{D}}(\widehat{t}_{yR}) = \sum_{k\in U}\sum_{l\in U} (\pi_{kl} - \pi_k\pi_l) \frac{(y_k - Bx_k)(y_l - Bx_l)}{\pi_k\pi_l}, \tag{10.29}$$

де

$$B = \frac{\sum\limits_{k\in U} y_k}{\sum\limits_{k\in U} x_k} = \frac{T_y}{T_x}.$$

Оцінка дисперсії оцінки $\widehat{t}_{yr}$

$$\widehat{\mathcal{D}}(\widehat{t}_{yR}) = \sum_{k\in s}\sum_{l\in s} \frac{(\pi_{kl} - \pi_k\pi_l)}{\pi_{kl}} \left(\frac{g_{ks}(y_k - \widehat{B}x_k)}{\pi_k}\right) \left(\frac{g_{ls}(y_l - \widehat{B}x_l)}{\pi_l}\right), \tag{10.30}$$

де для всіх $k \in s$

$$g_{ks} = \frac{\sum\limits_{k\in U} x_k}{\sum\limits_{k\in s} x_k/\pi_k} = \frac{T_x}{\widehat{t}_{x\pi}}. \tag{10.31}$$

Це твердження є наслідком застосування результатів підрозділу 10.2 до моделі відношення (10.27).

Якщо поділити вираз (10.28) на N, то отримаємо оцінку за відношенням для середнього значення досліджуваної характеристики генеральної сукупності:

$$\widehat{\overline{y}}_R = \overline{X}\, \frac{\sum\limits_{k\in s} y_k/\pi_k}{\sum\limits_{k\in s} x_k/\pi_k} = \overline{X}\, \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}}. \tag{10.32}$$

Щоб отримати формули для дисперсії та оцінки дисперсії оцінки $\widehat{\overline{y}}_R$, потрібно поділити відповідні вирази для оцінки $\widehat{t}_{yR}$ на N^2.

10.4. Вправи та питання для самоконтролю

10.1. Як саме використовується допоміжна інформація у методі оцінювання за різницею?

10.2. Чи є оцінка за регресією незміщеною? Від чого залежить ефективність цієї оцінки?

10.3. Чи пов'язані між собою оцінка за регресією та оцінка за відношенням?

10.4. Довести, що у випадку простого випадкового відбору без повернення з розміром вибірки $n = fN$ та з $y_k^0 = x_k$, дисперсія оцінки за різницею сумарного значення характеристики y має такий вигляд:

$$\mathcal{D}(\widehat{t}_{yD}) = N^2 \frac{1-f}{n} \left(S_y^2 + S_x^2 - 2S_{xy}\right). \qquad (10.33)$$

10.5. Нехай модель регресії з однією допоміжною змінною x має такий вигляд:

$$\begin{cases} E_\xi(y_k) = \beta x_k, \\ \mathcal{D}_\xi(y_k) = \sigma^2 x_k, \end{cases}$$

де значення $x_1, \ldots, x_N$ – додатні. Довести, що в цьому випадку

$$\widehat{B} = \frac{\sum\limits_{k\in s} y_k/\pi_k}{\sum\limits_{k\in s} x_k/\pi_k}.$$

10.6. Нехай модель регресії з однією допоміжною змінною x має такий вигляд:

$$\begin{cases} E_\xi(y_k) = \beta_1 + \beta_2 x_k, \\ \mathcal{D}_\xi(y_k) = \sigma^2. \end{cases}$$

Довести, що

$$\widehat{\mathbf{B}} = \begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\bar{y}} - \widehat{B}_2 \tilde{\bar{x}} \\ \widehat{B}_2 \end{pmatrix},$$

де

$$\widehat{B}_2 = \frac{\sum\limits_{k\in s} (x_k - \tilde{\bar{x}})(y_k - \tilde{\bar{y}})/\pi_k}{\sum\limits_{k\in s} (x_k - \tilde{\bar{x}})^2/\pi_k},$$

$$\tilde{\bar{x}} = \frac{\sum\limits_{k\in s} x_k/\pi_k}{\widehat{N}}, \quad \tilde{\bar{y}} = \frac{\sum\limits_{k\in s} y_k/\pi_k}{\widehat{N}}, \quad \widehat{N} = \sum_{k\in s} \frac{1}{\pi_k}.$$

10.7. Для випадку простого випадкового відбору без повернення вивести формулу для обчислення оцінки за відношенням сумарного значення досліджуваної характеристики генеральної сукупності, а також формули для обчислення наближеного значення дисперсії та оцінки дисперсії отриманої оцінки за відношенням.

10.8. Із генеральної сукупності розміру $N = 200$ вибіркових одиниць за допомогою простого випадкового відбору без повернення отримано вибірку розміру $n = 40$, за якою обчислено такі величини для досліджуваної змінної y та допоміжної змінної x:

$$\sum_{k\in s} y_k = 24874; \quad \sum_{k\in s} y_k^2 = 19136668;$$

$$\sum_{k\in s} x_k = 473; \quad \sum_{k\in s} x_k^2 = 6539; \quad \sum_{k\in s} x_k y_k = 348071.$$

Крім того, відомо, що $T_x = 2603$. Оцінити сумарне значення характеристики y за відношенням. Побудувати 95-відсотковий довірчий інтервал для сумарного значення характеристики y генеральної сукупності.

Розділ 11
Аналіз даних із пропусками

Відсутність відповіді (*невідповідь, пропуск*) для деяких елементів вибірки – це природна, проте небажана властивість вибіркових обстежень. Багато дослідників намагаються якомога швидше позбутись пропусків, щоб далі проводити обробку «повних» даних стандартними статистичними методами. Проте такий підхід може призвести до значних відмінностей статистичних висновків, зроблених за наявності пропусків та без них.

Розвиток комп'ютерних технологій дозволив статистикам застосовувати до аналізу даних методи з великою кількістю обчислень. Ці методи дають можливість отримувати кращі оцінки пропущених значень, вибіркових середніх, дисперсій та коваріацій [8,10]. Крім того, для отримання більш точних оцінок потрібно звертати увагу на механізм відбору елементів та причини виникнення пропусків.

Розрізняють два види невідповідей: *повна невідповідь* – коли повністю відсутня інформація для деякої вибіркової одиниці, та *часткова невідповідь* – коли при обстеженні вибіркової одиниці відсутня відповідь на одне або кілька окремих запитань.

Повні невідповіді виникають у таких випадках:

– інтерв'юер не має можливості провести обстеження;

– особа через хворобу не може відповісти на запитання;

– особа відмовляється відповідати.

У цих випадках інтерв'юер може зібрати допоміжну інформацію, яку можна використати для зменшення впливу невідповідей на оцінки.

Часткові невідповіді виникають найчастіше в результаті відмов відповідати на конкретні питання, пов'язані, наприклад, з прибутками, вживанням наркотиків тощо.

У сільськогосподарських обстеженнях та при обстеженнях диких тварин частіше користуються терміном «пропущені дані» (англ. `missing data`), але суть та наслідки залишаються такими самими.

11.1. Механізми породження пропусків

Знання або незнання механізму, що призводить до відсутності значень, є ключовим при виборі методу аналізу та інтерпретації результатів. Іноді цим механізмом керує статистик. Наприклад, ми можемо вважати, що вибірковому обстеженню властиві пропуски, оскільки значення частини змінних при обстеженні наявні в усіх елементів генеральної сукупності, а досліджувані змінні «пропущені» в елементів, що не були включені у вибірку. Тут механізм пропусків – процес побудови вибірки. Якщо елементи вибираються випадково, то механізм можна назвати «ігнорованим». Якщо правило отримання елементів вибірки не дотримується або для деяких вибіркових одиниць значення відсутні, то причини утворення пропусків стають менш зрозумілими. У цьому випадку аналіз залежить від припущень щодо механізму утворення пропусків, які потрібно явно обумовлювати.

Прикладом ситуації, коли механізм породження пропусків може бути некерованим, проте відомим статистику, є цензурування. Досліджувана змінна – це момент настання певної події (загибель тварини в експерименті, народження дитини, перегоряння лампочки), тобто, вибіркові дані – це моменти часу. Для деяких вибіркових одиниць момент настання події цензурований, бо подія ще не трапилась до закінчення експерименту. Якщо відома точка (час) цензурування, то ми маємо часткову інформацію про те, що момент настання неспостереженої події перевищує час цензурування. Таку інформацію потрібно враховувати в аналізі, щоб позбутися зміщення.

Розглянемо випадок двох змінних: допоміжної змінної X та досліджуваної змінної Y. Нехай пропуски можливі лише у змінній Y, а X спостерігається без пропусків. Механізм утворення пропусків для даних цієї структури корисно класифікувати відповідно до залежності ймовірності пропуску:

1) від Y та, можливо, від X;

2) від X, але не від Y;

3) ані від X, ані від Y.

Д. Б. Рубін [10] пропонує таку термінологію. Якщо виконується випадок 3), то ми кажемо, що пропущені дані відсутні випадково (англ. `missing at random`, MAR) і наявні дані присутні випадково (англ. `observed at random`, OAR), або в цілому дані відсутні абсолютно випадково (англ. `missing completely at random`, MCAR). У цьому випадку спостережувані значення Y утворюють випадкову підвибірку загальної вибірки. У випадку 2) ми кажемо, що пропущені дані відсутні випадково (MAR). Спостережувані значення Y необов'язково є випадковою підвибіркою отриманих даних, проте утворюють випадкову підвибірку в кожній підгрупі, що визначається значенням X. У випадку 1 дані не відповідають ані випадку MAR, ані OAR, тобто вони відсутні не випадково (англ. `missing not at random`, MNAR).

11.2. Огляд методів аналізу даних із пропусками

Усі методи обробки та аналізу вибіркових даних з пропусками можна наближено розподілити на такі чотири групи.

1. **Метод виключення некомплектних елементів**. За відсутності у деяких вибіркових одиниць значень яких-небудь змінних найпростішим прийомом є видалення таких некомплектних елементів та обробка даних без пропусків. Цей метод може застосовуватись при малій кількості пропусків. У інших випадках він може призвести до серйозних зміщень та, як правило, не є ефективним. Наприклад, при дослідженні доходів пропуски при низьких та високих доходах більш імовірні, ніж при середніх доходах. У результаті виключення некомплектних елементів частка спостережень із середнім доходом у вибірці збільшиться і може стати нерепрезентативною.

Інформацію, що міститься у виключених спостереженнях, можна використовувати для того, щоб дослідити, чи є повні спостереження випадковою підвибіркою вихідної вибірки. Наприклад, значна різниця між середніми окремої змінної Y_j для повних спостережень та для тих неповних спостережень, у яких присутня Y_j, вказує на систематичну відмінність некомплектних даних.

2. **Методи заповнення пропусків – імпутації** (англ. `imputation`). Пропуски заповнюють, і отримані «повні» дані аналізують звичайними методами. Як правило, використовуються такі процедури: заповнення з підбором, коли підставляються значення змінних інших об'єктів вибірки; заповнення середніми та заповнення за допомогою регресійної формули, отриманої для повних спостережень. Серед методів заповнення пропусків розрізняють процедури так званої «хот-дек імпутації» (англ. `hot-deck imputation`) та «колд-дек імпутації» (англ. `cold-deck imputation`). У випадку хот-дек імпутації пропущені значення заповнюють значеннями досліджуваної змінної, які вибирають (випадково або певним спеціальним методом) серед наявних у даному обстеженні. А у випадку колд-дек імпутації дані для заповнення пропусків беруть з інших джерел, наприклад, використовують результати попередніх обстежень або історичні дані. На відміну від процедур хот-дек імпутації, для заповнення за допомогою регресії використовуються «оцінені» співвідношення між різними змінними.

3. **Методи зважування**. Висновки за даними вибіркових обстежень з пропусками побудовані на так званих вагах дизайну. Ці ваги обернено пропорційні до ймовірностей включення [6]. Нехай y_i – значення змінної Y для i-го елемента генеральної сукупності. Тоді середнє значення досліджуваної характеристики генеральної сукупності можна оцінити величиною

$$\sum_{i\in s}\pi_i^{-1}y_i/\sum_{i\in s}\pi_i^{-1}, \tag{11.1}$$

де π_i – ймовірність включення i-го елемента у вибірку, π_i^{-1} – вага i-го елемента. Методи зважування змінюють (калібрують) ваги, щоб урахувати пропущенні значення. Оцінка (11.1) замінюється оцінкою

$$\sum(\pi_i\hat{p}_i)^{-1}y_i/\sum(\pi_i\hat{p}_i)^{-1}, \tag{11.2}$$

де суми беруться по тих елементах вибірки, в яких немає пропусків, а $\hat{p}_i$ – оцінка ймовірності відповіді, тобто присутності значення для i-го елемента (зазвичай це доля одиниць вибірки з прису-

тніми значеннями). Зважування пов'язане із заповненням середніми. Наприклад, якщо ваги дизайну є сталими в підгрупах вибірки, то заповнення пропусків у кожній підгрупі середніми підгрупи та зважування присутніх значень за допомогою їх долі в кожній підгрупі призводять до однакових оцінок середнього генеральної сукупності.

4. **Методи, що базуються на моделюванні.** Широкий клас методів базується на побудові моделі породження пропусків та розподілу даних. Висновки отримуються за допомогою функції вірогідності.

11.3. Методи заповнення пропусків

Перелічимо основні методи заповнення пропусків у вибіркових обстеженнях.

1) *Заповнення середніми* за присутніми значеннями у вибірці.

2) *Заповнення пропусків із підбором* полягає в заміні пропусків значеннями відповідних змінних, наявними у схожих вибіркових одиниць, які не мали пропусків.

3) *Заміна* застосовується на етапі збору даних при обстеженні. Цей метод полягає в заміні елемента з відсутніми відповідями на інший елемент, не включений у вибірку. Наприклад, якщо неможливе опитування домовласника, то можна опитати його сусіда, не включеного у список опитуваних. Однак, було б невірним розглядати отриману таким чином вибірку як повну, оскільки ті, хто дають відповіді, можуть систематично відрізнятися від тих, кого не вдалося опитати. Тому при аналізі слід розглядати цю заміну як заповнення певного виду.

4) *Заповнення без підбору* означає, що пропуск заповнюється деяким сталим значенням із зовнішнього джерела, наприклад, значенням попереднього спостереження з того самого обстеження. Отримані дані прийнято аналізувати як повні, тобто наслідки заповнення ігнорують.

5) *Заповнення за регресією* полягає в заповненні пропусків значеннями, отриманими за регресійною формулою для даного елемента. Зазвичай вона обчислюється за комплектними елементами.

Заповнення середніми можна розглядати як частковий випадок заповнення за регресією, якщо вважати предикаторами фіктивні змінні, що вказують на групу, в якій відбувається підстановка середніх.

6) *Стохастичне заповнення за регресією* полягає в заміні пропуску сумою значень, обчислених за регресією, та залишку, що відображає невизначеність передбачуваних значень. Природними поправками будуть залишки регресії, отримані для елементів без пропусків, що обираються випадково для кожного пропуску.

11.3.1. Заповнення середніми

Нехай y_i – значення досліджуваної характеристики Y для i -го елемента генеральної сукупності, $i = 1, ..., N$. Тоді найпростішою оцінкою пропущених значень y_i є середнє значення $\overline{y}$, обчислене за наявними значеннями змінної Y. Зрозуміло, що середнє спостережуваних та підставлених значень дорівнює $\overline{y}$, звичайній оцінці середнього за наявними даними.

Узагальнимо цей метод на випадок, коли спостережувані дані можна розділити на J груп, для яких відомо, що значення змінної Y у елементів всередині кожної групи схожі. Нехай y_{ij} – значення змінної Y для i-го елемента у групі j, $i = 1, ..., N_j$, $j = 1, ..., J$. При заповненні середніми для вибіркових елементів, що не дали відповідь, підставляється середнє $\overline{y}_j$ за тими m_j респондентами, що дали відповідь у j-й групі. Середнє генеральної сукупності $\overline{Y}$ можна оцінити середнім за присутніми та підставленими значеннями, а саме:

$$\sum_{j=1}^{J} n_j \hat{\overline{y}}_j / \sum_{j=1}^{J} n_j,$$

де $\hat{\overline{y}}_j$ – середнє присутніх та підставлених значень у j-й групі. Тепер

$$\hat{\overline{y}}_j = [m_j \overline{y}_j + (n_j - m_j)\overline{y}_j]/n_j = \overline{y}_j,$$

тому отримана оцінка $\overline{Y}$ – просто оцінка зі зважуванням груп.

Метод заповнення середніми реалізується просто, проте він має кілька небажаних властивостей. По-перше, правильні оцінки дисперсії $\overline{y}_j$ неможливо отримати за допомогою звичайних формул для дисперсії, що застосовуються для аналізу повних даних. Реальний розмір вибірки занижений через відсутність відповідей, тому звичайні формули призводять до занижених оцінок справжньої дисперсії. По-друге, величини, не лінійні за даними, такі, як дисперсія Y або кореляція між двома змінними, неможливо обґрунтовано оцінити стандартними методами для повних даних. По-третє, підстановка середніх суттєво змінює емпіричний розподіл значень Y, що є важливим при дослідженні розподілу Y за гістограмами або іншими графіками, що відображають дані. Усі ці недоліки спонукають шукати значення для пропусків, використовуючи методи заповнення типу підстановки з підбором. Звернемося тепер до цього методу.

11.3.2. Заповнення з підбором

Згідно з цим методом заповнення з підбором пропуски заповнюються значеннями, що були отримані для іншого подібного елемента вибірки. Припустимо, як і раніше, що отримана вибірка розміру n із N елементів генеральної сукупності, та тільки для m із n елементів вибірки зареєстровано значення досліджуваної змінної Y. Для простоти розмістимо елементи так, щоб m елементів, які не мають невідповідей, мали номери $1, ..., m$. При рівноймовірнісній схемі відбору середнє $\overline{Y}$ можна оцінити як середнє за отриманими та підставленими значеннями, що можна записати у вигляді

$$\overline{y} = [m\overline{y}_R + (n-m)\overline{y}_{NR}]/n, \qquad (11.3)$$

де $\overline{y}_R$ – середнє за отриманими відповідями; $\overline{y}_{NR} = \sum_{i=1}^{m} \frac{H_i y_i}{n-m}$ – середнє за заповненими значеннями; H_i – кратність, із якою y_i використовувалося для підстановки замість пропуску Y. Зауважимо, що $\sum_{i=1}^{m} H_i = n - m$, де $n - m$ – кількість елементів із пропусками.

Властивості $\overline{y}$ залежать від способу формування чисел $(H_1, ..., H_m)$. Найлегше отримати формули, якщо ймовірності підстановки наявних значень y_i, $i = m+1, ..., n$, однакові та дорівнюють $1/m$.

Інший метод генерування значень для заповнення пропусків – послідовний відбір, при якому всі елементи розташовують у певній послідовності та пропущене значення замінюється значенням Y для найближчого в цій послідовності попередника елемента, що дав відповідь.

Оцінки підстановки з підбором, які ми обговорювали досі, незміщені лише при загальному нереальному припущенні, що ймовірність відповіді не пов'язана зі значенням Y. Якщо є деяка додаткова інформація щодо елементів, що дають та не дають відповідь, то її можна використовувати для зменшення зміщення, що виникає через пропуски. Наступні два підходи заслуговують на увагу.

1) *Підстановка з підбором всередині груп.* Формуються групи, і пропуски в кожній групі заповнюються її ж значеннями. Середнє та дисперсію отриманих таким чином оцінок параметра $\overline{Y}$ знаходять, використовуючи наведені вище формули окремо всередині груп, а потім об'єднуючи отримані значення.

2) *Підбір найближчого сусіда.* Цей метод полягає у використанні метрики d для вимірювання відстані між елементами і виборі підстановки за елементом із наявним значенням, найближчим до елемента з пропуском. Наприклад, нехай $x_{i1}, ..., x_{iJ}$ – значення J основних змінних, виміряних у нормованих шкалах, у елемента i з пропуском y_i. Визначимо відстань

$$d(i,j) = \max_k |x_{ik} - x_{jk}|$$

між елементами i та j. Ми можемо обирати підстановку для y_i серед тих j-х елементів, для яких: 1) спостерігаються $y_j, x_{j1}, ..., x_{jJ}$ та 2) $d(i,j)$ не перевищує деякий поріг d_0. Число «кандидатів» – j-х елементів, що підходять, – можна обирати, змінюючи d_0. Оскільки значення, що підставляються, є достатньо складними

функціями присутніх ознак, властивості оцінок у таких процедурах підбору поки що недостатньо досліджені.

11.3.3. Заповнення за регресією

У цьому методі пропущені значення заповнюються за допомогою регресії змінних із пропусками на інші змінні.

Нехай $\mathbf{X}$ – $n \times p$–матриця, в якій i-й рядок $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ містить значення змінних, що спостерігаються без пропусків $(i = \overline{1, n})$, та $\mathbf{y} = (y_1, ..., y_n)'$ – вектор значень змінної, в якій є пропуски. Вважаємо, що значення $y_1, ..., y_m$ присутні, а $y_{m+1}, ..., y_n$ – пропущені. Припустимо, що виконується така лінійна модель:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e},$$

де $\mathbf{e} = (e_1, ..., e_n)'$, e_i – незалежні однаково розподілені випадкові величини з нульовим середнім та однаковою дисперсією σ^2, а β – параметр розмірності p. Застосуємо метод найменших квадратів для знаходження оцінок параметрів β та σ^2 за повними спостереженнями. Якщо $(\mathbf{X}'\mathbf{X})$ має повний ранг, то оцінка найменших квадратів параметра β обчислюється так:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}).$$

У протилежному випадку оцінка найменших квадратів параметра β невизначена.

Якщо $(\mathbf{X}'\mathbf{X})$ невироджена, то $\hat{\beta}$ – незміщена оцінка параметра β з найменшою дисперсією. Якщо e_i розподілені нормально, то $\hat{\beta}$ – оцінка максимальної вірогідності параметра β, розподілена нормально із середнім β та дисперсією $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Найкращою незміщеною оцінкою дисперсії σ^2 є

$$\hat{\sigma}^2 = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{n - p},$$

де $\hat{y}_i = \mathbf{x}_i\hat{\beta}$. Якщо e_i нормальні, то $(n - p)\hat{\sigma}^2/\sigma^2$ має розподіл хі-квадрат із $n - p$ ступенями вільності.

Найкраща незміщена оцінка коваріаційної матриці $\hat{\beta} - \beta$ дорівнює

$$\mathbf{V} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Якщо e_i розподілені нормально, то $(\hat{\beta}_i - \beta_i)/\sqrt{v_{ii}}$ (де v_{ii} – i-й діагональний елемент матриці $\mathbf{V}$) має розподіл Стьюдента з $n - p$ ступенями вільності.

Отже, щоб отримати заповнені дані, потрібно покласти

$$\hat{y}_i = \begin{cases} y_i, i = 1, ..., m, \\ \mathbf{x}_i\hat{\beta}, i = m + 1, ..., n. \end{cases}$$

С. Ф. Бак [13] показав, що середні значення, обчислені за присутніми та підставленими значеннями – правильні оцінки середніх лише за умови MCAR. Проте вибіркова коваріаційна матриця, обчислена за заповненими даними, занижує значення дисперсій та коваріацій, хоч і не так сильно, як при заповненні середніми. Це пов'язано з відсутністю відхилень підставлених значень від регресійної прямої.

11.4. Оцінювання вибіркової дисперсії за наявності пропусків

Розглянемо побудову оцінок вибіркової дисперсії, які включають додатковий член для врахування пропусків.

Важливо підкреслити, що для багатьох застосувань питання зміщень через пропуски часто більш важливе, ніж оцінка дисперсії. У цьому розділі вважається, що введені поправки, що виправляють зміщення, пов'язані із пропусками.

Обчислення правильних оцінок дисперсії для складних вибіркових дизайнів, що часто використовуються на практиці, – непросте завдання навіть при повних даних. Тому були розвинені наближені методи, що можуть застосовуватись для більшості вибіркових дизайнів. Простота цих методів зумовлена тим, що обчислення зводяться до розрахунку величин для множини вибіркових одиниць, що називаються скінченними кластерами (СК). СК – найбільша вибіркова одиниця, що вилучається з вибірки.

Оцінювання дисперсії базується на такій лемі.

Лема 11.1. *Нехай* $\hat{\theta}_1, ..., \hat{\theta}_k$ *– некорельовані випадкові величини, які мають спільне середнє* μ, *та*

$$\bar{\theta} = \sum_{j=1}^{k} \frac{\hat{\theta}_j}{k},$$

$$\widehat{\mathcal{D}}(\bar{\theta}) = \sum_{j=1}^{k} \frac{(\hat{\theta}_j - \bar{\theta})^2}{k(k-1)}.$$

Тоді

1) $\bar{\theta}$ *– незміщена оцінка параметра* μ,

2) $\widehat{\mathcal{D}}(\bar{\theta})$ *– незміщена оцінка дисперсії* $\bar{\theta}$.

Доведення. 1) $E(\bar{\theta}) = \sum\limits_{j=1}^{k} E(\hat{\theta}_j)/k = \mu$.

2) Помітимо, що

$$\sum_{j=1}^{k}(\hat{\theta}_j - \mu)^2 = \sum_{j=1}^{k}(\hat{\theta}_j - \bar{\theta})^2 + k(\bar{\theta} - \mu)^2.$$

Звідси

$$E(\sum_{j=1}^{k}(\hat{\theta}_j - \bar{\theta})^2) - k(k-1)\mathcal{D}(\bar{\theta}) =$$

$$= \sum_{j=1}^{k}\mathcal{D}(\hat{\theta}_j) - k\mathcal{D}(\bar{\theta}) - k(k-1)\mathcal{D}(\bar{\theta}) =$$

$$= \sum_{j=1}^{k}\mathcal{D}(\hat{\theta}_j) - k^2\mathcal{D}(\bar{\theta}). \qquad (11.4)$$

Проте

$$k^2\mathcal{D}(\bar{\theta}) = \mathcal{D}(\sum_{j=1}^{k}\hat{\theta}_j) = \sum_{j=1}^{k}\mathcal{D}(\hat{\theta}_j),$$

оскільки оцінки $\hat{\theta}_j$ некорельовані. Отже, вираз (11.4) дорівнює нулю, що доводить 2). □

Цю лему можна безпосередньо застосовувати до оцінок у вибіркових дизайнах із поверненням. Розглянемо такий приклад.

П р и к л а д 11.1. Розглянемо генеральну сукупність, яка складається із K скінченних кластерів. Нехай вибірковий дизайн задає одержання k кластерів шляхом простого випадкового відбору з поверненням. Оцінимо сумарне значення генеральної сукупності

$$T = \sum_{j=1}^{K} T_j,$$

де T_j – сумарне значення змінної Y у j-ому кластері.

За Горвіцем–Томпсоном

$$\widehat{t}_\pi = \frac{1}{k}\sum_{j=1}^{k}\hat{t}_j/\pi_j,$$

де сумування ведеться за вибраними СК, $\hat{t}_j$ – незміщена оцінка сумарного значення T_j; π_j – ймовірність обрання j-го СК. Тоді маємо: 1) $\widehat{t}_\pi$ та усі $\hat{t}_j/\pi_j, j = 1, ..., k$, – незміщені оцінки параметра T; та 2) оцінки $\hat{t}_j/\pi_j, j = 1, ..., k$, – некорельовані. Звідси за попередньою лемою незміщена оцінка дисперсії оцінки $\widehat{t}_\pi$ обчислюється так:

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = \sum_{j=1}^{k}\frac{(\hat{t}_j/\pi_j - \widehat{t}_\pi)^2}{k(k-1)}. \qquad (11.5)$$

◊

Припустимо, що у вибірці є пропуски і ми отримуємо оцінки $\hat{t}_j$ сумарних значень у СК за допомогою одного з методів обробки пропусків, що наводилися вище. Тоді ми також можемо застосовувати оцінку (11.5) для оцінки дисперсії, якщо: 1) оцінки $\hat{t}_j$ незміщені; 2) поправки на заповнення або зважування обираються всередині кожного СК незалежно, щоб оцінки $\hat{t}_j$ залишалися некорельованими.

На практиці рідко відбувається вибір із поверненням. Коли проводиться простий випадковий відбір без повернення, оцінки СК від'ємно корельовані. Д. Б. Рубін показав, що оцінки виду (11.5), що базуються на лемі, завищують дисперсію. Якщо ж увести поправку на скінченність генеральної сукупності $(1 - k/K)$, то отримана оцінка дисперсії буде заниженою. Для незміщеної оцінки потрібна інформація про другий та наступні етапи обстеження. Таким чином, для побудови простих оцінок дисперсії, що базуються на СК, потрібно, щоб доля вибраних СК була мала, що дасть змогу нехтувати зміщенням, обумовленим вибором без повернення. У практичних дослідженнях така ситуація трапляється досить часто.

Більшість вибіркових дизайнів містять стратифікацію при виборі СК. Припускаючи, що доля СК у кожній страті мала, за допомогою оцінок за СК можна вивести обґрунтовані оцінки дисперсії. Припустимо, що є H страт. Нехай $\hat{t}_{hj}$ – незміщена оцінка суми T_{hj} для j-го кластера у страті h, $h = 1, ..., H$, $j = 1, ..., K_h$.

Можна оцінити T величиною

$$\hat{t} = \sum_{h=1}^{H} \sum_{j=1}^{k_h} \frac{\hat{t}_{hj}}{\pi_{hj}} = \sum_{h=1}^{H} \hat{t}_h,$$

де сума обчислюється за H стратами та k_h елементами, включеними у вибірку зі страти h, π_{hj} – ймовірність вибору j-го СК у страті h, а $\hat{t}_h$ – оцінка суми у страті h. Оцінкою дисперсії $\hat{t}$ є

$$\widehat{\mathcal{D}}(\hat{t}) = \sum_{h=1}^{H} \sum_{j=1}^{k_h} \frac{(\hat{t}_{hj}\pi_{hj} - \hat{t}_h)^2}{k_h(k_h - 1)}.$$

Зокрема, при виборі двох СК із кожної страти оцінкою дисперсії є

$$\widehat{\mathcal{D}}(\hat{t}) = \sum_{h=1}^{H} \sum_{j=1}^{k_h} \frac{(\hat{t}_{h1}\pi_{h1} - \hat{t}_{h2}\pi_{h2})^2}{4}.$$

Умови, за яких можна отримати ці оцінки за заповненими даними, такі самі, як і у випадку простого випадкового відбору: підстановки потрібно виконувати незалежно в кожному СК.

11.5. Аналіз даних із пропусками за допомогою функції вірогідності

11.5.1. Повні дані

Розглянемо спочатку випадок повних даних. Позначимо дані через Y. Y може бути скаляром, вектором або матрицею залежно від контексту. Припустимо, що дані породжуються згідно з моделлю, яка описується функцією розподілу або щільністю $f(Y \mid \theta)$, що залежить від скалярного або векторного параметра θ.

Функцією вірогідності $L(\theta \mid Y)$ називається будь-яка функція від θ, що пропорційна до $f(Y \mid \theta)$, при фіксованому Y.

Логарифмічною функцією вірогідності $l(\theta \mid Y)$ називають натуральний логарифм функції вірогідності $L(\theta \mid Y)$.

Приклад 11.2. *Одновимірна нормальна вибірка.* Сумісна щільність n незалежних однаково розподілених спостережень $Y = (y_1, \ldots, y_n)'$ із нормального розподілу з середнім μ та дисперсією σ^2 має вигляд

$$f(Y \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu)^2}{\sigma^2}\right).$$

При фіксованому Y логарифм функції вірогідності

$$l(\mu, \sigma^2 \mid Y) = \ln(f(Y \mid \mu, \sigma^2)),$$

або, відкидаючи адитивну сталу, маємо

$$l(\mu, \sigma^2 \mid Y) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu)^2}{\sigma^2}, \qquad (11.6)$$

що є функцією $\theta = (\mu, \sigma^2)'$ при фіксованих спостережуваних даних Y. ◊

Приклад 11.3. *Експоненціальна вибірка.* Сумісна щільність незалежних однаково розподілених спостережень із експоненціального розподілу має вигляд

$$f(Y \mid \theta) = \theta^{-n} \exp\left(-\sum_{i=1}^{n}\frac{y_i}{\theta}\right).$$

Знайдемо логарифм функції вірогідності, що розглядається як функція θ при фіксованих спостережуваних даних Y:

$$l(\theta \mid Y) = \ln\left\{\theta^{-n}\exp\left(-\sum_{i=1}^{n}\frac{y_i}{\theta}\right)\right\} = -n\ln\theta - \sum_{i=1}^{n}\frac{y_i}{\theta}. \qquad (11.7)$$

$\Diamond$

Оцінкою максимальної вірогідності (ОМВ) $\hat{\theta}$ називають значення параметра θ, яке максимізує функцію вірогідності $L(\theta \mid Y)$, або, що еквівалентно, логарифм функції вірогідності $l(\theta \mid Y)$.

Сформульоване твердження допускає можливість існування більш ніж однієї ОМВ. Проте в багатьох важливих моделях ОМВ єдина, до того ж функція вірогідності диференційована та обмежена зверху. Для таких випадків ОМВ можна знайти, якщо прирівняти похідну функції вірогідності по θ до нуля та розв'язати отримане рівняння відносно θ. Рівняння

$$S(\theta \mid Y) \equiv \frac{\partial l(\theta \mid Y)}{\partial\theta} = 0$$

називають *рівнянням максимальної вірогідності*, а похідну логарифма функції вірогідності $S(\theta \mid Y)$ – *функцією внеску* (впливу) вибірки.

Приклад 11.4. *Експоненціальна вибірка.* Логарифм вірогідності для вибірки з експоненціального розподілу визначається виразом (11.13). Диференціюванням за θ отримуємо рівняння максимальної вірогідності:

$$-\frac{n}{\theta} + \sum_{i=1}^{n}\frac{y_i}{\theta^2} = 0.$$

Розв'язавши його відносно θ, отримаємо ОМВ: $\hat{\theta} = \bar{y} = \sum_{i=1}^{n} y_i/n$ – середнє вибірки Y. $\Diamond$

Приклад 11.5. *Одновимірна нормальна вибірка.* Використовуючи формулу (11.6), запишемо логарифм вірогідності для вибірки розміру n із нормального розподілу так:

$$l(\mu, \sigma^2 \mid Y) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2} =$$

$$= -\frac{n}{2}\ln\sigma^2 - \frac{1}{2}\frac{n(\bar{y}-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{n\hat{S}^2}{\sigma^2},$$

де $\hat{S}^2 = n^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ – вибіркова дисперсія. Диференціюючи за θ та прирівнюючи похідну до нуля, отримаємо

$$(\bar{y}-\mu)/\sigma^2 = 0,$$

що призводить до $\hat{\mu} = \bar{y}$. Диференціюючи за σ^2 та прирівнюючи похідну до нуля, маємо

$$-\frac{n}{2\sigma^2} + \frac{n(\bar{y}-\mu)^2}{2\sigma^4} + \frac{n\hat{S}^2}{2\sigma^2} = 0,$$

що призводить до $\hat{\sigma} = \hat{S}^2$, оскільки $\hat{\mu} = \bar{y}$. Отже, ми отримали ОМВ:

$$\hat{\mu} = \bar{y},\ \hat{\sigma} = \hat{S}^2.$$

$\Diamond$

11.5.2. Оцінювання методом максимальної вірогідності за неповними даними

Позначимо через Y дані, що спостерігалися б за відсутності пропусків. Тоді $Y = (Y_{obs}, Y_{mis})$, де Y_{obs} позначає наявні, а Y_{mis} – пропущені дані. Позначимо через $f(Y \mid \theta) = f(Y_{obs}, Y_{mis} \mid \theta)$ ймовірність або щільність сумісного розподілу Y_{obs} та Y_{mis}. Інтегруванням за пропущеними даними Y_{mis} отримаємо щільність імовірності

$$f(Y_{obs} \mid \theta) = \int f(Y_{obs}, Y_{mis} \mid \theta)dY_{mis}.$$

Визначимо функцію вірогідності від θ на основі Y_{obs} без урахування механізму породження пропусків як будь-яку функцію від θ, пропорційну $f(Y_{obs} \mid \theta)$:

$$L(\theta \mid Y_{obs}) \sim f(Y_{obs} \mid \theta). \tag{11.8}$$

У більш загальному випадку ми включаємо до моделі розподіл змінної, що вказує на наявність кожного елемента з Y. Індикатором пропуску назвемо величину, що набуває значення 1, якщо даний елемент спостерігається, та 0, якщо він не спостерігається. Наприклад, якщо $Y = (Y_{ij}) \in n \times K$-матриця n спостережень над K-вимірною змінною, індикатор пропуску визначається таким чином:

$$R_i = \begin{cases} 1, \text{ якщо значення } y_i \text{ спостерігається,} \\ 0, \text{ якщо значення } y_i \text{ пропущене.} \end{cases}$$

У цій моделі R розглядається як випадкова змінна та визначається сумісний розподіл R та Y. Щільність цього розподілу можна задати як добуток щільності розподілу Y та щільності умовного розподілу R при фіксованому Y, тобто

$$f(Y, R \mid \theta, \psi) = f(Y \mid \theta) f(R \mid Y, \psi).$$

Будемо називати умовний розподіл R при заданому Y, що залежить від невідомого параметра ψ, розподілом пропусків. У деяких випадках цей розподіл повністю відомий, і параметризація за допомогою ψ не потрібна.

Фактично спостережувані дані складаються зі значень змінних (Y_{obs}, R). Розподіл спостережуваних даних можна отримати, якщо проінтегрувати сумісну щільність $Y = (Y_{obs}, Y_{mis})$ за R та Y_{mis}:

$$f(Y_{obs}, R \mid \theta, \psi) = \int f(Y_{obs}, Y_{mis} \mid \theta) f(R, Y_{obs}, Y_{mis}, \psi) dY_{mis}. \tag{11.9}$$

Функція вірогідності від θ та ψ – будь-яка функція, пропорційна функції (11.9):

$$L(\theta, \psi \mid Y_{obs}, R) \sim f(Y_{obs}, R \mid \theta, \psi). \tag{11.10}$$

Тепер постає питання: коли слід будувати висновки відносно θ за допомогою функції вірогідності $L(\theta, \psi \mid Y_{obs}, R)$, яка визначається співвідношенням (11.10), а коли – за допомогою більш простого виразу $L(\theta \mid Y_{obs})$ з (11.8), в якому механізм породження пропусків ігнорується. Помітимо, що при незалежності розподілу пропусків від пропущених значень Y_{mis}, тобто при

$$f(R \mid Y_{obs}, Y_{mis}, \psi) = f(R \mid Y_{obs}, \psi) \tag{11.11}$$

з формули (11.9) випливає, що

$$f(Y_{obs}, R \mid \theta, \psi) = f(R, Y_{obs}, \psi) \int f(Y_{obs}, Y_{mis} \mid \theta) dY_{mis} =$$

$$= f(R, Y_{obs}, \psi) f(Y_{obs} \mid \theta).$$

У багатьох важливих застосуваннях параметри θ та ψ розділені в тому сенсі, що сумісний параметричний простір (θ, ψ) є добутком параметричних просторів для θ та ψ. Якщо θ та ψ розділені, то висновки відносно θ, що базуються на функції вірогідності $L(\theta, \psi \mid Y_{obs}, R)$, будуть збігатися з висновками, що базуються на $L(\theta \mid Y_{obs})$. Тому, якщо вірне рівняння (11.11), то механізмом породження пропусків можна знехтувати – отримані функції вірогідностей пропорційні.

За наведеним вище означенням дані відсутні випадково (MAR), коли виконується співвідношення (11.11). Для практики важливим є той факт, що для ефективного застосування методів, що базуються на функції вірогідності, потрібно виконання лише умови MAR, а не більш жорсткої умови MCAR.

Приклад 11.6. *Неповна експоненціальна вибірка.* Припустимо, ми маємо неповну одновимірну вибірку з експоненціального розподілу, в якій присутні значення $Y_{obs} = (y_1, ..., y_m)'$ та відсутні $Y_{mis} = (Y_{m+1}, ..., y_n)'$. Отже, як у попередньому прикладі

$$f(Y \mid \theta) = \theta^{-n} \exp\left(-\sum_{i=1}^{n} \frac{y_i}{\theta}\right).$$

Функція вірогідності, коли механізм пропусків ігнорується, пропорційна щільності Y_{obs} при заданому θ, яка визначається виразом:

$$f(Y \mid \theta) = \theta^{-m} \exp\left(-\sum_{i=1}^{m} \frac{y_i}{\theta}\right). \qquad (11.12)$$

Для цього прикладу $R = (R_1, ..., R_n)'$, де $R_i = 1$, $i = 1, ..., m$, та $R_i = 0$, $i = m+1, ..., n$.

Припустимо, що кожний елемент спостерігається з імовірністю ψ, так що виконується (11.11). Тоді

$$f(R \mid Y, \psi) = \psi^m (1-\psi)^{n-m},$$

$$f(Y_{obs}, R \mid \theta, \psi) = \psi^m (1-\psi)^{n-m} \theta^{-m} \exp\left(-\sum_{i=1}^{m} \frac{y_i}{\theta}\right).$$

Якщо θ та ψ розділяються, то висновки відносно θ можна робити за $f(Y_{obs} \mid \theta)$, ігноруючи механізм пропусків. Зокрема, ОМВ параметра θ дорівнює $\sum_{i=1}^{n} y_i/m$ – середньому за наявними значеннями Y.

Тепер припустимо, що пропуски утворюються внаслідок цензурування в деякій відомій точці c, так що наявні лише значення, менші за c. Тоді

$$f(R \mid Y, \psi) = \prod_{i=1}^{n} f(R_i \mid y_i, \psi),$$

де

$$f(R_i \mid y_i, \psi) = \begin{cases} 1, \; R_i = 1 \text{ та } y_i < c \text{ або } R_i = 0 \text{ та } y_i > c; \\ 0 - \text{у протилежному випадку.} \end{cases}$$

Отже,

$$f(Y_{obs}, R \mid \theta) = \prod_{i=1}^{m} f(y_i, R_i \mid \theta) \prod_{i=m+1}^{n} f(R_i \mid \theta) =$$

$$= \prod_{i=1}^{m} f(y_i \mid \theta) f(R_i \mid y_i) \prod_{i=m+1}^{n} P(y_i > c \mid \theta) =$$

$$= \theta^{-m} \exp\left(-\sum_{i=1}^{m} \frac{y_i}{\theta}\right) \exp\left(-\frac{(n-m)c}{\theta}\right), \qquad (11.13)$$

оскільки $P(y_i > c \mid \theta) = \exp(-c/\theta)$ згідно з властивостями експоненціального розподілу. У цьому випадку механізм породження пропусків не можна ігнорувати, та точна функція вірогідності за формулою (11.13) відрізняється від (11.12). Максимізація (11.12) дає ОМВ $\hat{\theta} = (\sum_{i=1}^{n} y_i + (n-m)c)/m$, яка більша порівняно з раніше знайденою оцінкою $\sum_{i=1}^{n} y_i/m$. Додатна поправка до вибіркового середнього пов'язана з цензуруванням неспостережуваних значень. ◊

11.6. *EM*-алгоритм

EM-алгоритм (англ. `expectation-maximization algorithm`) – це загальний ітеративний алгоритм для оцінювання методом максимальної вірогідності в задачах з неповними даними. Цей алгоритм корисно застосовувати у випадку, коли важко знайти явний розв'язок рівняння максимальної вірогідності. У цьому алгоритмі реалізована така ідея обробки неповних даних:

1) заповнення пропусків оцінками пропущених значень;

2) оцінювання параметрів;

3) повторне оцінювання пропущених значень, при цьому оцінки параметрів вважаються точними;

4) повторне оцінювання параметрів і так далі.

Кожна ітерація *EM*-алгоритму складається з кроку *E* (обчислення математичного сподівання) та кроку *M* (максимізація).

Опис кроку M досить простий: «Проводити оцінювання параметра θ методом максимальної вірогідності так, ніби немає пропусків». Отже на цьому кроці використовуються ті самі обчислювальні методи, що і при максимізації $l(\theta \mid Y)$.

На кроці E знаходять умовне математичне сподівання пропущених даних при фіксованих спостережуваних даних та оцінках параметрів, а потім замінюють пропущені дані знайденими очікуваними значеннями. Причому, для реалізації алгоритму необов'язково заповнювати пропущені дані, достатньо оцінити деякі функції пропущених даних, що входять у логарифм функції вірогідності для повних даних $l(\theta \mid Y)$.

Ще однією перевагою EM-алгоритму є його надійна збіжність. Тобто, в деяких нестрогих припущеннях кожна ітерація збільшує логарифм функції вірогідності $l(\theta \mid Y)$, та якщо $l(\theta \mid Y)$ обмежений, то послідовність $l(\theta^{(t)} \mid Y)$ збігається до стаціонарного значення $l(\theta \mid Y)$. Також відомо, що якщо послідовність $l(\theta^{(t)} \mid Y)$ збігається, то вона збігається до локального максимуму або точки перегину $l(\theta \mid Y)$. Недолік EM-алгоритму полягає в тому, що швидкість збіжності може бути дуже низькою, якщо пропущено багато даних.

Точніше, нехай $\theta^{(t)}$ – поточна оцінка параметра θ. На кроці E EM-алгоритму знаходять очікуваний логарифм функції вірогідності (математичне сподівання логарифма функції вірогідності) за умови $\theta = \theta^{(t)}$:

$$Q(\theta \mid \theta^{(t)}) = \int l(\theta \mid Y) f(Y_{mis} \mid Y_{obs}, \theta = \theta^{(t)}) dY_{mis}.$$

На кроці M EM-алгоритму визначають $\theta^{(t+1)}$, максимізуючи цей очікуваний логарифм функції вірогідності:

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) \qquad \text{для всіх } \theta.$$

Приклад 11.7. *Одновимірні нормальні дані.* Припустимо, що y_i незалежні однаково нормально розподілені з середнім μ та дисперсією σ^2. Причому y_i для $i = 1, ..., m$ спостерігаються, а y_i для $i = m+1, ..., n$ пропущені, та виконується припущення MAR. Математичне сподівання значення кожного пропуску y_i при заданих

Y_{obs} та $\theta = (\mu, \sigma^2)$ дорівнює μ. При цьому, згідно з формулою (11.6), логарифм функції вірогідності виражається через достатні статистики $\sum_{i=1}^{n} y_i$ та $\sum_{i=1}^{n} y_i^2$. Отже, на кроці E алгоритму отримаємо значення

$$E\left(\sum_{i=1}^{n} y_i \Big| \theta^{(t)}, Y_{obs}\right) = \sum_{i=1}^{m} y_i + (n-m)\mu^{(t)}, \tag{11.14}$$

$$E\left(\sum_{i=1}^{n} y_i^2 \Big| \theta^{(t)}, Y_{obs}\right) = \sum_{i=1}^{m} y_i^2 + (n-m)\left((\mu^{(t)})^2 + (\sigma^{(t)})^2\right) \tag{11.15}$$

для поточних оцінок параметрів $\theta^{(t)} = (\mu^{(t)}, (\sigma^{(t)})^2)$. З останньої рівності бачимо, що проста підстановка $\mu^{(t)}$ замість пропусків $y_{m+1}, ..., y_n$ призвела б до відсутності члена $(n-m)(\sigma^{(t)})^2$.

Для повних даних ОМВ параметра μ дорівнює $\sum_{i=1}^{n} y_i/n$, а ОМВ параметра σ^2 – це $\sum_{i=1}^{n} y_i^2/n - (\sum_{i=1}^{n} y_i/n)^2$. На кроці M використовуються ті самі поточні математичні сподівання достатніх статистик, обчислених на кроці E. Отже, на кроці M обчислюються

$$\mu^{(t+1)} = E\left(\sum_{i=1}^{n} y_i \mid \theta^{(t)}, Y_{obs}\right)/n \tag{11.16}$$

та

$$(\sigma^{(t+1)})^2 = E\left(\sum_{i=1}^{n} y_i^2 \mid \theta^{(t)}, Y_{obs}\right)/n - (\mu^{(t+1)})^2. \tag{11.17}$$

Покладаючи в рівняннях (11.14) – (11.17) $\mu^{(t)} = \mu^{(t+1)} = \hat{\mu}$ та $\sigma^{(t)} = \sigma^{(t+1)} = \hat{\sigma}$, отримаємо тотожності. Отже, алгоритм збігається до

$$\hat{\mu} = \sum_{i=1}^{m} y_i/m$$

та

$$\hat{\sigma}^2 = \sum_{i=1}^{m} y_i^2/m - \hat{\mu}^2,$$

тобто до ОМВ параметрів μ та σ^2 за Y_{obs} за умови виконання припущення MAR. $\Diamond$

11.7. Вправи та питання для самоконтролю

11.1. Як впливають пропуски у вибіркових обстеженнях на оцінки параметрів генеральної сукупності?

11.2. Які методи аналізу даних з пропусками ви можете назвати?

11.3. У чому полягає метод заповнення пропущених даних за регресією?

11.4. За якого припущення щодо механізму породження пропусків можна ефективно використовувати методи аналізу неповних даних, які ґрунтуються на застосуванні функції вірогідності?

11.5. Охарактеризувати функції кроків E та M EM-алгоритму.

11.6. Знайти ОМВ коефіцієнта варіації μ/σ одновимірної нормальної вибірки.

11.7. Генеральна сукупність, яка містить $N = 10000$ елементів, поділена на дві страти обсягом $N_1 = 1000$ та $N_2 = 9000$ елементів відповідно. З кожної страти відібрано по 500 елементів за допомогою простого випадкового відбору без повернення. 250 елементів вибірки з першої страти та 50 елементів вибірки з другої страти обстежити не вдалося (тобто, є 250 невідповідей у вибірці з першої страти та 50 – у вибірці з другої страти). Оцінити ймовірність відповіді для кожної страти окремо та для популяції в цілому. Знайти ваги дизайну для даного обстеження. Як зміняться ваги дизайну, якщо врахувати пропущені значення? Записати формулу для обчислення оцінки середнього значення досліджуваної характеристики генеральної сукупності, використовуючи наявні числові дані та калібровані ваги дизайну.

Додаток 1. Таблиця випадкових чисел

Числа, наведені у таблиці Д1, можна використовувати для побудови ймовірнісної вибірки.

Таблиця Д1. **Випадкові числа**

4924	2621	2514	4455	9711	1976	0308	3212	8638
7680	0568	8778	3996	0998	8625	4506	6684	7793
5970	4902	1544	2810	8221	4022	3616	7954	6257
2987	1686	2788	9285	5947	9673	2614	5966	9133
2646	1809	7469	6628	5080	5025	3906	7979	6269
4848	4229	1076	8969	8756	2836	2893	7430	3620
5497	8722	6911	0794	9419	3160	8038	2607	5299
1670	4811	4211	5073	2348	0956	5165	9942	1715
1485	4611	4765	7903	0638	1406	5404	5430	8715
7794	2374	6830	9832	8038	1727	7501	3898	7937
5643	2263	0444	0889	0698	2109	1195	2140	2585
8879	5094	9197	6246	4311	3016	9587	7750	4391
1715	3085	4729	2761	5988	8949	6692	9764	2889
6591	3228	5133	1429	9187	6325	6240	0089	1000
0031	2817	7924	7565	5523	5999	1482	1377	5698
6722	2565	9574	2048	0178	9170	2517	1930	2078
3659	1905	0394	5019	4839	2854	1262	1434	6583
5514	5849	6951	4852	7391	4533	2545	2608	7408
8940	2013	6147	1136	3906	6087	6054	7899	4834
0996	1176	9037	2954	2611	6859	2751	3540	4568
3155	1666	2782	4984	7135	7289	4662	3274	4902
7686	0141	8403	7507	5481	2868	7362	7259	3453
7385	1820	9399	4384	8736	0097	1745	0924	5357
8742	3039	0170	0927	2970	7634	0094	3087	1702
7189	8624	7943	4307	9255	4939	8986	2376	6746
7429	3532	4075	6833	5550	2816	7272	4624	2848
7535	0249	1106	9795	0864	1380	6000	8889	1481
3005	9164	7251	6231	7949	7927	9971	8322	2980
9844	0573	0733	8485	6749	2572	5178	7846	8996
1096	9787	5211	4342	8500	1435	2986	1057	4450
7960	6896	6454	9658	2578	1956	0563	4190	8844

Додаток 2. Нормальний розподіл

У таблиці Д2 наведено значення функції $\Phi(t)$ нормального розподілу з параметрами (0; 1):

- для заданих $t \geq 0$ табульовані значення функції

$$N_{0;\,1}(t) = \Phi(t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{t} \exp\left\{-\frac{s^2}{2}\right\} ds.$$

- для $t < 0$ значення функції $\Phi(t)$ отримуються з рівності

$$\Phi(t) = 1 - \Phi(-t).$$

Значення $N_{a;\,\sigma^2}(x)$ функції нормального розподілу з параметрами a і σ^2 обчислюється за значеннями табульованої функції $N_{0;\,1}(t) = \Phi(t)$ нормального розподілу $N_{0;\,1}$:

$$N_{a;\,\sigma^2}(x) = N_{0;\,1}\left(\frac{x-a}{\sigma}\right) = \Phi\left(\frac{x-a}{\sigma}\right).$$

Таблиця Д2 допускає лінійну інтерполяцію.

Таблиця Д 2. **Значення функції $\Phi(t)$**

t	0	1	2	3	4	5	6	7	8	9
0,0	,5000	,5040	,5080	,5120	,5160	,5199	,5239	,5279	,5319	,5359
0,1	,5398	,5438	,5478	,5517	,5557	,5596	,5636	,5675	,5714	,5753
0,2	,5793	,5832	,5871	,5910	,5948	,5987	,6026	,6064	,6103	,6141
0,3	,6179	,6217	,6255	,6293	,6331	,6368	,6406	,6443	,6480	,6517
0,4	,6554	,6591	,6628	,6664	,6700	,6736	,6772	,6808	,6844	,6879
0,5	,6915	,6950	,6985	,7019	,7054	,7088	,7123	,7157	,7190	,7224
0,6	,7257	,7291	,7324	,7357	,7389	,7422	,7454	,7486	,7517	,7549
0,7	,7580	,7611	,7642	,7673	,7703	,7734	,7764	,7794	,7823	,7852
0,8	,7881	,7910	,7939	,7967	,7995	,8023	,8051	,8078	,8106	,8133
0,9	,8159	,8186	,8212	,8238	,8264	,8289	,8315	,8340	,8365	,8389
1,0	,8413	,8438	,8461	,8485	,8508	,8531	,8554	,8577	,8599	,8621
1,1	,8643	,8665	,8686	,8708	,8729	,8749	,8770	,8790	,8810	,8830
1,2	,8849	,8869	,8888	,8907	,8925	,8944	,8962	,8980	,8997	,9015
1,3	,9032	,9049	,9066	,9082	,9099	,9115	,9131	,9147	,9162	,9177
1,4	,9192	,9207	,9222	,9236	,9251	,9265	,9279	,9292	,9306	,9319
1,5	,9332	,9345	,9357	,9370	,9382	,9394	,9406	,9418	,9429	,9441
1,6	,9452	,9463	,9474	,9484	,9495	,9505	,9515	,9525	,9535	,9545
1,7	,9554	,9564	,9573	,9582	,9591	,9599	,9608	,9616	,9625	,9633
1,8	,9641	,9649	,9656	,9664	,9671	,9678	,9686	,9693	,9699	,9706
1,9	,9713	,9719	,9726	,9732	,9738	,9744	,9750	,9756	,9761	,9767
2,0	,9772	,9778	,9783	,9788	,9793	,9798	,9803	,9808	,9812	,9817
2,1	,9821	,9826	,9830	,9834	,9838	,9842	,9846	,9850	,9854	,9857
2,2	,9861	,9864	,9868	,9871	,9875	,9878	,9881	,9884	,9887	,9890
2,3	,9893	,9896	,9898	,9900	,9904	,9906	,9909	,9911	,9913	,9916
2,4	,9918	,9920	,9922	,9925	,9927	,9929	,9931	,9932	,9934	,9936
2,5	,9938	,9940	,9941	,9943	,9945	,9946	,9948	,9949	,9951	,9952
2,6	,9953	,9955	,9956	,9957	,9959	,9960	,9961	,9962	,9963	,9964
2,7	,9965	,9966	,9967	,9968	,9969	,9970	,9971	,9972	,9973	,9974
2,8	,9974	,9975	,9976	,9977	,9977	,9978	,9979	,9979	,9980	,9981
2,9	,9981	,9982	,9982	,9983	,9984	9984	,9985	,9985	,9986	,9986
t	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9
$\Phi(t)$	,9987	,9990	,9993	,9995	,9997	,9998	,9998	,9999	,9999	,1000

Покажчик термінів

Список основних позначень

$E(\widehat{\theta})$ – математичне сподівання оцінки $\widehat{\theta}$;
$\mathcal{D}(\widehat{\theta})$ – дисперсія оцінки $\widehat{\theta}$;
$\widehat{\mathcal{D}}(\widehat{\theta})$ – оцінка дисперсії оцінки $\widehat{\theta}$;
N – кількість елементів популяції;
n, n_s – розмір вибірки;
$p(\cdot)$ – вибірковий дизайн;
$S_y^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \overline{Y})^2$ – дисперсія змінної y для популяції;
$\widehat{S}_y^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \overline{y})^2$ – вибіркова дисперсія змінної y;
s – вибірка;
$T = \sum_{k\in U} y_k$ – сумарне значення змінної y для популяції;
$t = \sum_{k\in s} y_k$ – сумарне значення змінної y за вибіркою;
$\widehat{t}_\pi$ – оцінка Горвіца–Томпсона сумарного значення T;
U – генеральна сукупність, популяція;
U_d – підсукупність;
U_h – страта;
x_k – значення допоміжної змінної x для елемента k;
$\overline{Y} = \frac{1}{N}\sum_{k\in U} y_k$ – середнє значення змінної y для популяції;
$\overline{y} = \frac{1}{n}\sum_{k\in s} y_k$ – вибіркове середнє змінної y;
y_k, z_k – значення змінних y та z для елемента k;
$\widehat{\overline{y}}_\pi$ – оцінка Горвіца–Томпсона середнього значення $\overline{Y}$;
θ – параметр генеральної сукупності;
π_k – ймовірність включення першого порядку;
π_{kl} – ймовірність включення другого порядку.

Список скорочень

ВБ – відбір Бернуллі;
ВВО – вторинна вибіркова одиниця;
ВП – відбір Пуассона;
ДВЕ – двостадійний відбір елементів;
ДКВ – двостадійний кластерний відбір;
ОКВ – одностадійний кластерний відбір;
ПВВбП – простий випадковий відбір без повернення;
ПВВзП – простий випадковий відбір з поверненням;
ПВО – первинна вибіркова одиниця;
СВ – систематичний відбір;
СТВ – стратифікований відбір;
СТПВВ – стратифікований простий випадковий відбір.

Список рекомендованої літератури

1. Вибіркове спостереження. Термінологічний словник / О. О. Васєчко, О. І. Черняк, Є. М. Жуйкова та ін. – К. : ІВЦ ДКС України, 2004.

2. *Зінченко, Н. М.* Аналітичні моделі та методи соціології / Н. М. Зінченко, А. Я. Оленко. – К. : ВПЦ «Київський ун-т», 2000.

3. *Карташов, М. В.* Імовірність, процеси, статистика / М. В. Карташов. – К. : ВПЦ «Київський ун-т», 2007.

4. *Майборода, Р. Є.* Регресія: лінійні моделі / Р. Є. Майборода. – К. : ВПЦ «Київський ун-т», 2007.

5. *Пархоменко, В. М.* Методи вибіркових обстежень / В. М. Пархоменко. – К. : ТВіМС, 2001.

6. *Саріогло, В. Г.* Проблеми статистичного зважування вибіркових даних / В. Г. Саріогло. – К. : ІВЦ Держкомстату України, 2005.

7. *Черняк, О. І.* Техніка вибіркових досліджень / О. І. Черняк. – К. : МІВВЦ, 2001.

8. *Щербіна, А. М.* Пропуски у вибіркових обстеженнях / А. М. Щербіна // Кваліфікаційна робота на здобуття ступеня бакалавра математики. – К. : Київський національний ун-т імені Тараса Шевченка, 2007.

9. *Кокрен, У.* Методы выборочного исследования / У. Кокрен. – М. : Статистика, 1976.

10. *Литтл, Р. Дж. А.* Статистический анализ данных с пропусками / Р. Дж. А. Литтл, Д. Б. Рубин. – М. : Финансы и статистика, 1991.

11. *Ardilly, P.* Sampling Methods. Exercises and Solutions / P. Ardilly, Y. Tillé. – Springer Science+Business Media Inc., 2006.

12. *Brewer, K. R. W.* Sampling with Unequal Probabilities / K. R. W. Brewer, M. Hanif. – Springer-Verlag, 1983.

13. *Buck, S. F.* A method of estimation of missing values in multivariate data suitable for use with an electronic computer / S. F. Buck // Journal of the Royal Statistical Society, Series B, 1960. – Vol. 22.

14. *Fan, C. N.* Development of sampling plans by using sequential (item by item) techiques and digital computers / C. N. Fan, M. E. Muller, I. Rezucha // Journal of the American Statistical Association, 1962. – Vol. 57.

15. *Hájek, J.* Limiting distributions in simple random sampling from a finite population / J. Hájek // Publ. Math. Inst. Hung. Acad. Sci., 1960. – Vol. 5.

16. *Isaki, C. T.* Survey design under the regression superpopulation model / C. T. Isaki, W. A. Fuller // Journal of the American Statistical Association, 1982. – Vol. 77.

17. *Kish, L.* Survey Sampling / L. Kish. – Wiley, 1995.

18. *Lohr, S.* Sampling: design and analysis / S. Lohr. – New York : Duxbury Press, 1999.

19. *McLeod, A. L.* A convenient algorithm for drawing a simple random sample / A. L. McLeod, D. R. Bellhouse // Applied Statistics, 1983. – Vol. 32, No. 2.

20. *Särndal, C.-E.* Model Assisted Survey Sampling / C.-E. Särndal, B. Swensson, J. Wretman. – New York : Springer-Verlag, 1992.

21. *Sunter, A. B.* Response burden, sample rotation and classification renewal in economic surveys / A. B. Sunter // International Statistical Review, 1977. – Vol. 45.

22. *Sunter, A. B.* List sequential sampling with equal or unequal probabilities without replacement / A. B. Sunter // Applied Statistics, 1977. – Vol. 26, No. 3.

Навчальне видання

ВАСИЛИК Ольга Іванівна
ЯКОВЕНКО Тетяна Олександрівна

ЛЕКЦІЇ З ТЕОРІЇ І МЕТОДІВ ВИБІРКОВИХ ОБСТЕЖЕНЬ

Навчальний посібник

Редактор *Л. В. Магда*

Оригінал-макет виготовлено Видавничо-поліграфічним центром "Київський університет"

**Підписано до друку 07.09.10. Формат 60x84 1/16. Вид. № 148. Гарнітура Times.
Папір офсетний. Друк офсетний. Наклад 350.
Ум. друк. арк. 12,09. Обл.-вид. арк. 13. Зам. № 210-5338.**

**Видавничо-поліграфічний центр "Київський університет"
01601, Київ, б-р Т. Шевченка, 14, кімн. 43
☎ (38044) 239 3222; (38044) 239 3172; факс (38044) 239 31 28
Свідоцтво внесено до Державного реєстру ДК № 1103 від 31.10.02.**