КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
імені ТАРАСА ШЕВЧЕНКА

О.І.Василик, Т. О. Яневич

ЗБІРНИК ЗАДАЧ З ТЕОРІЇ ВИБІРКОВИХ ОБСТЕЖЕНЬ

Видавничо-поліграфічний центр
«Київський університет»
2022

Збірник задач з теорії вибіркових обстежень: Навчальний посібник / О.І.Василик, Т. О. Яневич– К.: ВПЦ «Київський університет», 2022. – 120 с.

Рецензенти

доктор фіз.-мат. наук, доцент І. В. Розора
кандидат фіз.-мат. наук, доцент О. А. Лагода

Затверджено Вченою Радою
механіко–математичного факультету
07.02.2022 р.

Зміст

Передмова

Даний збірник задач є посібником до нормативного курсу «Вибіркові обстеження», що викладається студентам магістратури спеціальностей «Прикладна та теоретична статистика» та «Актуарна та фінансова математика» на механіко-математичному факультеті Київського національного університету імені Тараса Шевченка. Він доповнює підручник «Лекції по теорії і методам вибіркових обстежень», що був виданий авторами у 2010 році.Останній містить теоретичний матеріал. Цей посібник призначений для використання на практичних заняттях, а також для самостійної роботи студентів. Багато ідей для цього задачника було почерпнуто під час спілкування з колегами із інших університетів, що приймали участь у конференціях, воркшопах та літніх школах Балтійсько-скандинаво-української мережі зі статистики обстежень.

Курс «Вибіркові обстеження» викладається на механіко-математичному факультеті КНУ близько двадцяти 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 [6] та О. І. Черняка [8]. З перетворенням цього курсу в нормативний та збільшенням кількості аудиторних годин, виникла потреба в додатковому матеріалі. Впродовж років ми накопичували базу корисних практичних задач із багатьох джерел, основними з яких є [9, 11, 18, 20, 24]. У результаті виникла ідея підготувати навчальний посібник, який зібрав всі ці задачі разом. Оскільки останні два-три роки навчання в університетах було переважно в дистанційному форматі і студенти часто потребували самостійно розбиратись в задачах, тому було вирішено для деяких типових задач включити розв'язки. Для інших задач в посібнику вказані або відповіді, або вказівки, що дають змогу студентам якісно опрацьовувати практичну складову курсу.

Посібник складається з десяти основних розділів, що містять базовий теоретичний матеріал та задачі по відповідним темах. Крім того, в розділі 11 містяться формулювання типових лабо-

раторних робіт, що використовують різні віртуальні середовища: область Стефенс, Острови та Статвілідж. Ці середовища дають змогу краще зрозуміти принципи і певні особливості теорії вибіркових обстежень. Зокрема, студенти вчаться здобувати та опрацьовувати необроблені статистичні дані та оцінювати якість отриманих результатів не знаючи, якими є шукані параметри насправді. Розділ 12 містить вказівки, відповіді та розв'язки до задач із основних розділів. Завершується посібник списком рекомендованої літератури.

Перший розділ містить задачі, що сприяє опрацюванню студентами основних понять, які використовуються в теорії вибіркових обстежень: вибірковий дизайн, ймовірності включення першого та другого порядку, π-оцінка Горвіца–Томпсона та оцінка її дисперсії 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 необхідний теоретичний матеріал та задачі на використання та обчислення оцінки Горвіца-Томсона при різних методах відбору. Зокрема, другий розділ присвячено простому випадковому відбору без повернення (ПВВбП). В ньому включено задачі на оцінювання різних лінійних параметрів генеральної сукупності (сумарне, середнє, пропорція) та підсукупностей; на побудову довірчих інтервалів; на визначення необхідного розміру вибірки у випадку застосування цього відбору. У третьому розділі містяться задачі на опрацювання методів оцінювання для відбору Бернуллі. Тут також розглянуто поняття дизайн-ефекту, за допомогою якого визначається ефективність довільного методу відбору у порівнянні з ПВВбП. Четвертий розділ присвячений задачам, які стосуються систематичного відбору. У п'ятому розділі розглядається відбір з поверненням, опрацьовується оцінка Хансена–Гурвіца та порівнюється із оцінкою Горвіца-Томсона при відборі із повереннням. А в шостому розділі включено завдання на опрацювання деяких методів нерівноймовірнісного відбору без повернення. У сьомому розділі детально вивчається стратифі-

кований відбір. Значна увага приділяється задачам підрахунку та порівнянню точності при використанні пропорційного, Неймановського та оптимального розміщень при простому стратифікованому відборі. В задачах пропонується порівняти ефективність цього вибіркового дизайну при різних розміщеннях та при використанні ПВВбП. Задачі на опрацювання одностадійного, двостадійного та багатостадійного кластерного відбору представлені у восьмому розділі, де в основному розглядається випадок застосування ПВВбП на окремих стадіях одно- та двостадійного відбору. Наступні два розділи присвячені більш складним задачам оцінювання: в дев'ятому розділі розглядається задачі оцінювання лінійних та нелінійних функцій від сумарних значень кількох досліджуваних змінних, а в десятому – задачі оцінювання за різницею, за регресією та за відношенням. Студентам пропонується не тільки навчитись рахувати необхідні оцінки, а також порівнювати та аналізувати ефективність використання цих методів оцінювання.

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 посібника висловлюють щиру подяку рецензентам посібника – доктору фіз.-мат. наук, доценту І. В. Розорі та кандидату фіз.-мат. наук, доценту О. А. Лагоді 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 теплим словом хочемо згадати професора Гуннара Кулдорфа з університету міста Умео (Швеція), який був ідейним натхненником розроблення та постійного розвитку курсу «Вибіркові обстеження» на механіко-математичному факультеті нашого університету.

Розділ 1

Вибірковий дизайн. Оцінка Горвіца-Томпсона

1.1. Основний теоретичний матеріал

Вибірковим дизайном, що відповідає деякому ймовірнісному методу відбору називається функція, що задає ймовірнісний розподіл на множині $\mathfrak{F}$ всіх можливих вибірок (підмножин), що можна утворити із елементів генеральної сукупності:

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

Випадкова величина, що є індикатором потрапляння елемента k у вибірку:

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

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

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

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

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

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

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

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

Оцінкою Горвіца–Томпсона або просто π-*оцінкою* для сумарного значення T досліджуваної характеристики y генеральної сукупності називається:

$$\widehat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}, \tag{1.1}$$

що є незміщеною оцінкою параметра T.

Дисперсія оцінки Горвіца-Томпсона обчислюється за формулою

$$\mathcal{D}(\widehat{t}_\pi) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in U} \sum_{l \in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l. \tag{1.2}$$

Якщо $\pi_{kl} > 0$ для всіх $k, l \in U$, то незміщеною оцінкою дисперсії $\mathcal{D}(\widehat{t}_\pi)$ буде статистика

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in s} \sum_{l \in s} \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l. \tag{1.3}$$

1.2. Задачі

Задача 1.1. 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 без повернення з N осіб відбираються n осіб, а потім визначають домогосподарства, до яких належать відібрані особи.

Підрахуйте ймовірність включення першого порядку у вибірку для домогосподарства, що складається з m осіб ($m < n$). Отримайте наближений вираз для цієї ймовірності при $m = 1, 2, 3$, припускаючи, що n і N є досить великими та $\frac{n}{N} \approx \frac{n-1}{N-1} \approx \frac{n-2}{N-2} = f = 0,1$.

Задача 1.2. Розглянемо генеральну сукупність $U = \{1, 2, 3\}$ з вибірковим дизайном $p(\cdot)$, при якому

$$p(\{1,2\}) = \frac{1}{2}, \quad p(\{1,3\}) = \frac{1}{3}, \quad p(\{2,3\}) = \frac{1}{6}.$$

Знайдіть ймовірності включення першого та другого порядків та коваріаційну матрицю індикаторів включення $\Delta = \{\Delta_{kl} = \mathrm{cov}(I_k, I_l)\}_{k,l\in U}$.

Задача 1.3. Нехай коваріаційна матриця індикаторів включення $\Delta = \{\Delta_{kl} = \mathrm{cov}(I_k, I_l)\}_{k,l\in U}$ при деякому дизайні $p(\cdot)$ має вигляд

$$\Delta = \frac{2}{9} \times \begin{pmatrix} 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 \end{pmatrix}.$$

1) Чи є $p(\cdot)$ дизайном з фіксованим розміром вибірки?
2) Чи задовольняє цей дизайн умову Єйтса–Гранді–Сена?
3) Обчисліть ймовірності включення першого порядку для дизайну $p(\cdot)$, якщо $\pi_1 = \pi_2 < \pi_3 = \pi_4 = \pi_5$.
4) Запишіть матрицю ймовірностей включення другого порядку $\Pi = \{\pi_{kl}\}_{k,l\in U}$.
6) Проаналізуйте отримані результати. Випишіть вибірковий дизайн, що відповідає заданій коваріаційній матриці. Який метод відбору можна використати для того, щоб його отримати?

Задача 1.4. 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) Обчисліть всі ймовірності включення першого та другого порядку.
2) Знайдіть $\mathrm{E}n_s$ та $\mathcal{D}n_s$, використовуючи означення математичного сподівання та дисперсії дискретної випадкової величини.
3) Знайдіть $\mathrm{E}n_s$ та $\mathcal{D}n_s$, використовуючи формули, що пов'язують ці величини із ймовірностями включення першого та другого порядків.

Задача 1.5. Нехай для генеральної сукупності та вибіркового дизайну з задачі 1.4 характеристика y має такі значення: $y_1 = 0$,

$y_2 = 1$, $y_3 = 1$, $y_4 = 0$ та $y_5 = 1$. Знайдіть:
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 $\widehat{\mathcal{D}}(\widehat{t}_\pi)$.

Задача 1.6. 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. Обчисліть яким в середньому буде розмір вибірки $\mathrm{E}n_s$ та якою буде дисперсія цієї величини $\mathcal{D}n_s$.

Розділ 2

Простий випадковий відбір без повернення

2.1. Основний теоретичний матеріал

Простий випадковий відбір без повернення.
Простим випадковим відбором без поверення (ПВВбП) називається відбір, що реалізується таким чином: з генеральної сукупності розміру N відбираються n ($n \leq N$) елементів з рівними ймовірностями та без повернення. В результаті отримаємо ймовірнісну вибірку розміру n. Всі такі вибірки будуть мати однакові ймовірності бути отриманими, а вибірковий дизайн матиме вигляд:

$$p(s) = \begin{cases} \frac{1}{C_N^n}, & \text{якщо розмір вибірки дорівнює } n; \\ 0, & \text{у протилежному випадку.} \end{cases} \tag{2.1}$$

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

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

Оцінювання сумарного та середнього значень при ПВВбП.

При простому випадковому відборі без повернення оцінки Горвіца-Томпсона сумарного та середнього значень досліджуваної характеристики y генеральної сукупності набувають такого вигляду

$$\widehat{t}_\pi = \frac{N}{n} \sum_{k \in s} y_k = \frac{N}{n} t, \text{ де } t = \sum_{k \in s} y_k \text{– вибіркове сумарне значення;}$$

$$\widehat{\overline{y}}_\pi = \frac{1}{N} \widehat{t}_\pi = \frac{t}{n} = \frac{1}{n} \sum_{k \in s} y_k = \overline{y}, \text{ де } \overline{y} \text{ – вибіркове середнє.}$$

Дисперсії цих оцінок дорівнюють:

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

Незміщені оцінки дисперсій обчислюються так:

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

де $S^2 = \frac{1}{N-1}\sum\limits_{k\in U}(y_k - \overline{Y})^2$ – дисперсія елементів у генеральній сукупності, $\widehat{S}^2 = \frac{1}{n-1}\sum\limits_{k\in s}(y_k - \overline{y})^2$ – дисперсія елементів у вибірці, $f = \frac{n}{N}$ – частка відбору.

Оцінювання пропорції при ПВВбП.

Розглянемо деяку підсукупність $U_d \subset U$, що цікавить дослідника (наприклад, сукупність безробітних). Нехай N_d – це кількість елементів, що належать даній підсукупності, тоді $P_d = \frac{N_d}{N}$ – це пропорція елементів в генеральній сукупності U, що належать підсукупності U_d.

При використанні простого випадкового відбору без повернення оцінка Горвіца-Томпсона пропорції P_d елементів, що належать заданій підсукупності U_d, має вигляд:

$$\begin{aligned} \widehat{P}_d &= p_d; \\ \mathcal{D}_{\text{ПВВбП}}\left(\widehat{P}_d\right) &= \frac{N-n}{N-1}\cdot\frac{P_d\cdot Q_d}{n}; \\ \widehat{\mathcal{D}}_{\text{ПВВбП}}\left(\widehat{P}_d\right) &= (1-f)\cdot\frac{p_d q_d}{n-1}, \end{aligned}$$

де $Q_d = 1 - P_d$, $p_d = \frac{n_d}{n}$ – пропорція елементів у вибірці, що належать підсукупності U_d, $q_d = 1 - p_d$.

Побудова довірчих інтервалів.

Якщо $\widehat{\theta}$ – це оцінка невідомого параметра θ, то довірчий інтервал для θ з рівнем довіри, що приблизно дорівнює $1-\alpha$, будують так:

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

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

Визначення необхідного розміру вибірки.

Для визначення необхідного розміру вибірки при оцінюванні параметра θ із заданою *допустимою похибкою* e та *рівнем довіри* $1-\alpha$ можна керуватись співвідношенням:

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

При оцінюванні середнього $\overline{Y}$ за допомогою π-оцінки при ПВВбП необхідний розмір вибірки повинен задовольняти нерівність

$$n \ge \frac{z^2_{1-\alpha/2}S^2}{e^2 + \frac{z^2_{1-\alpha/2}S^2}{N}}, \tag{2.2}$$

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

У разі задання точності відносною похибкою $\widetilde{e}$ необхідний розмір вибірки знаходять із нерівності

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

У цьому випадку значення n при ПВВбП та оцінюванні параметра $\overline{Y}$ за допомогою оцінки Горвіца-Томсона буде залежати вже не від дисперсії в генеральній сукупності, а від коефіцієнта варіації $CV_y = \sqrt{S^2}/\overline{Y}$:

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

2.2. Задачі

Задача 2.1. Метою обстеження є оцінка площі фермерських угідь у деякому районі Київської області. У цьому районі зареєстровано 2 200 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:

$$\sum_{k\in s} y_k = 3\,000\,\text{га} \quad \text{та} \quad \sum_{k\in s} y_k^2 = 156\,000\,\text{га}^2.$$

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.

Задача 2.2. Щоб реалізувати вибірку, що відповідає вибірковому дизайну ПВВбП (2.1), можна використати деякі альтернативні схеми відбору.

Отримайте вибірку розміру 4 з генеральної сукупності розміру 10 за допомогою *схем відбору 1, 2, 3*, що наведені нижче.

***Схема відбору 1** [14]*

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 ми ставимо у відповідність випадкове значення ε_k. Тоді ми перебираємо по-черзі елементи генеральної сукупності та вирішуємо включати елемент чи ні з ймовірністю, що дорівнює відношенню кількості елементів, що залишилось включити у вибірку до кількості елементів, що залишилось переглянути. А саме:

крок 1 *Якщо $\varepsilon_1 < n/N$, тоді елемент 1 відбирається та ні – в іншому випадку.*

Нехай n_k – це кількість елементів відібраних у вибірку із перших $k-1$ елементів зі списку.

крок k *Якщо $\varepsilon_k < \frac{n-n_k}{N-k+1}$, то k – відбирається, та ні – в іншому випадку.*

Процедура закінчується коли $n_k = n$.

В випадку, коли розмір генеральної сукупності невідомий, для отримання вибірки, що має вибірковий дизайн ПВВбП можна використати таку схему відбору.

***Схема відбору 2** [19]*

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 ймовірністю $\frac{n}{n+1}$. Якщо елемент $n+1$ потрапив таким чином у вибірку, тоді з неї викидається один елемент вибраний випадковим чином з однаковою для всіх імовірністю. В 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 реалізованою навіть без відомого наперед розміру генеральної сукупності N. Формування списку елементів сукупності може бути одночасним з процедурою відбору.

***Схема відбору 3** [22, 23]*

Основна перевага даної вибіркової схеми полягає у можливості одночасно вибрати декілька простих випадкових вибірок, що не перетинаються. Такі вибірки називаються від'ємно-координованими.

Схема полягає у наступному:

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 (2.1).

Задача 2.3. Типове соціологічне опитування громадської думки робиться на основі вибірки, що складається приблизно з 2 000 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 5 000 незалежних вибірок. Яка ймовірність того, що Ви *не* потрапите в жодну з цих вибірок?

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

Задача 2.4. Уряд періодично збирає інформацію щодо стану сільського господарства в країні. Розглядається генеральна сукупність із 1469 об'єднаних територіальних громад (ОТГ) України. Досліджується змінна y, що містить інформацію про площу відведених під фермерство земель в кожній ОТГ. Для 19 з 1469 ОТГ ця інформація невідома (тобто змінна містить пропущені значення). Яка ймовірність того, що при ПВВбП у вибірці розміру 300 не буде жодного пропущеного значення?

Задача 2.5. Знайти нерівність, яку повинен задовольняти необхідний розмір вибірки n якщо потрібно оцінити параметр *сумарне* T за допомогою оцінки Горвіца-Томсона $\widehat{t_\pi}$ із заданою допустимою похибкою e та рівнем довіри $1-\alpha$ при ПВВбП із генеральної сукупності, що складається із N елементів та має дисперсію S^2.

Задача 2.6. Проводиться обстеження працівників підприємства з метою визначення пропорції тих, хто хворіє на професійну хворобу. На підприємстві працює 4 000 працівників. Із зовнішніх джерел відомо, що зазвичай на подібних підприємствах професійну хворобу мають 3 з 10 працівників. При обстеженні планується вико-

ристання 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?

Задача 2.7. 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 $CV(\widehat{P}) = \frac{\sqrt{\mathcal{D}(\widehat{P})}}{P}$ не перевищував 5 %? Проаналізувати отриманий результат.

Задача 2.8. При обстеженні жителів деякого великого міста будуть досліджуватись дві величини:

P_1 – частка тих, хто має доступ до мережі інтернет;

P_2 – частка тих, хто має стаціонарний телефон.

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

Задача 2.9. В університеті працює 800 викладачів та спеціалістів-дослідників. В комп'ютерній системі можна подивитись список цих викладачів та їх рецензовані публікації. Для кожного викладача зафіксовано кількість таких рецензованих публікацій станом на початок 2021 року. Щоб зібрати інформацію про кількість рецензованих робіт, дослідник повинен отримати кожен запис окремо. Тому було вирішено зробити обстеження на основі вибірки. В табличці наведено кількості публікацій для вибірки з 50 викладачів (наприклад, 4 викладачі мають по одній рецензованій публікації). Для відбору викладачі у вибірку використовувався ПВВбП:

Рецензовані публікації	0	1	2	3	4	5	6	7	8	9	10
Викладачі	28	4	3	4	4	2	1	0	2	1	1

1. Зобразіть дані за допомогою гістограми. Опишіть, яку форму та особливості мають ці дані.

2. Оцініть, яка середня кількість публікацій приходиться на одного викладача за допомогою оцінки Горвіца-Томспона $\widehat{\bar{y}}_\pi = \bar{y}$. Оцініть коефіцієнт варіації для вашої оцінки, тобто підрахуйте наступну величина: $cv(\widehat{\bar{y}}_\pi) = \frac{\sqrt{\hat{\mathcal{D}}(\widehat{\bar{y}}_\pi)}}{\widehat{\bar{y}}_\pi}$.

3. На ваш погляд, чи буде $\bar{y}$ з пункту 2 нормально розподіленою величиною? Якщо так/ні, то чому?

4. Оцініть пропорцію викладачів, які не мають публікацій (тобто мають 0 рецензованих публікацій) і запишіть 95% довірчий інтервал для цієї оцінки.

5. Скільки потрібно обстежити викладачів при ПВВбП, щоб отримати оцінку для пропорції тих, хто не мають публікацій, із точністю 10% та надійністю 0.95?

Задача 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.

Розділ 3

Відбір Бернуллі

3.1. Основний теоретичний матеріал

Відбір Бернуллі (ВБ) базується на схемі Бернуллі та реалізується наступним чином. Розглянемо впорядковану в список генеральну сукупність, $k = 1, 2, ..., N$. 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.

Вибірковий дизайн при ВБ

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

де n_s – це розмір вибірки.

Ймовірності включення

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

Оцінка Горвіца-Томпсона при відборі Бернуллі.

При вибірковому дизайні Бернуллі оцінки Горвіца–Томпсона для сумарного T, середнього $\overline{Y}$ та пропорції P_d набувають вигляду

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

з дисперсіями

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

та оцінками для дисперсій

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

Дизайн-ефектом називається величина

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

Величина $deff$ визначає ефективність використання іншого вибіркового дизайну $p(\cdot)$ порівняно з простим випадковим відбором, якщо використовується оцінка Горвіца-Томсона $\widehat{t}_\pi$.

3.2. Задачі

Задача 3.1. 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.

Задача 3.2. 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 $n = n_s$.

Задача 3.3. 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 $n_{s_d} = 350$. Потрібно:

1. оцінити відсоток тих домогосподарств, що перебувають за межею бідності в цьому населеному пункті;

2. побудувати 95-% довірчий інтервал для цієї оцінки;

3. оцінити дизайн-ефект такого відбору Бернуллі, тобто, при однакових середній розмірах вибірки $n = \pi N$ підрахувати:

$$\widehat{deff}\left(\text{ВБ}, \widehat{P}_{d\pi}\right) := \frac{\widehat{\mathcal{D}}_{\text{ВБ}}(\widehat{P}_{d\pi})}{\widehat{\mathcal{D}}_{\text{ПВВбП}}(\widehat{P}_{d\pi})} = \frac{\frac{1-\pi}{N^2\pi^2} n_{s_d}}{\left(1 - \frac{n}{N}\right) \frac{p_d q_d}{n-1}},$$

де $p_d = \frac{n_{s_d}}{n}$, $q_d = 1 - p_d$.

Задача 3.4. Для виявлення сумарної кількості собак, яким була зроблена вакцинація від сказу, проводиться вибіркове обстеження в деякому регіоні України. Відомо, що в цьому регіоні є 120 ветеринарних пунктів, де проводиться вакцинація собак. Характеристикою y_k, що вивчається, є кількість собак, яким були зроблені щеплення в ветпункті k впродовж року. Для обстеження планується використання відбору Бернуллі. Відомо, що попереднього року $\sum_{k \in U} y_k \approx 2\,400$ та $\sum_{k \in U} y_k^2 \approx 50\,000$. Якою має бути середня кількість ветпунктів, що підлягають обстеженню, при такому відборі та застосуванні

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

2) альтернативної оцінки $\widehat{t}_{\text{альт}} = \frac{N}{n_s}\sum\limits_{k\in s} y_k = \frac{n_s}{n}\widehat{t}_\pi$, для якої $\mathcal{D}_{\text{ВБ}}(\widehat{t}_{\text{альт}}) \approx \mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)$,

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

Порівняйте отримані результати. Які висновки можна тут зробити?

Задача 3.5. Для генеральної сукупності U розміру $N = 10\ 000$ коефіцієнт варіації $CV_U = \sqrt{S^2}/\overline{Y} = 0,534$. 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.

Порівняйте отримані значення. Які висновки можна тут зробити?

Розділ 4

Систематичний відбір

4.1. Основний теоретичний матеріал

Систематичний відбір (СВ) полягає у виборі елементів у вибірку із деяким постійним кроком. 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* , 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, $r+a$, $r+2a$, $\ldots$ а вибірковий дизайн має вигляд

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

Ймовірності включення першого та другого порядків при СВ:

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

При застосуванні СВ важливим є контроль розміру n можливих вибірок. Нехай $N = an + c, \quad 0 \leq c < a$. Якщо $c = 0$, то всі a можливих вибірок мають однаковий розмір n; якщо $c > 0$, то розмір вибірки дорівнює $n+1$ при $r \leq c$ та n при $r > c$. Для отримання систематичних вибірок однакового розміру можна скористатись такими методами систематичного відбору.

Метод дробового вибіркового інтервалу. Тут допускається, 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

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

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

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

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

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

Оцінка Горвіца-Томсона. При систематичному відборі з вибірковим інтервалом a, π-оцінка для сумарного T має вигляд

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

з дисперсією

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

де

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

Вибірковий дизайн при систематичному відборі *не є вимірним*, тому оцінити дисперсію π-оцінки стандартним чином не можна. Є такі два способи оцінювання дисперсії при СВ.

Спосіб 1. Зміщена оцінка для дисперсії. Якщо є підстави вважати, що систематичний відбір такий же ефективний як і ПВВбП, тоді за оцінку дисперсії $\mathcal{D}_{\text{СВ}}\left(\widehat{t}_\pi\right)$ можемо вибрати

$$\widehat{\mathcal{D}}_{\text{СВ1}} = N^2 \frac{1-f}{n} \widehat{S}_r^2, \quad \text{де} \quad \widehat{S}_r^2 = \frac{1}{n-1} \sum_{k \in s_r} (y_k - \bar{y}_{s_r})^2. \tag{4.2}$$

В інших випадках ця оцінка може бути зміщеною, тому використовувати її треба із обережністю.

Спосіб 2. Модифікація систематичного відбору. 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 $\frac{n}{m}$.

Нехай $\frac{n}{m}$ та $a = \frac{N}{n}$ – цілі числа. Випадковим чином вибираєть ся 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:

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

При такій модифікації дизайн буде вже вимірним і ми можемо скористатись незміщеною оцінкою для дисперсії Горвіца–Томпсона для простого одностадійного кластерного відбору, в який він посуті перетворюється при такій модифікації:

$$\widehat{\mathcal{D}}_{\mathrm{CBm}}\left(\widehat{t}_\pi\right) = (ma)^2 \frac{1 - \frac{1}{a}}{m} S_m^2 = ma(a-1) S_m^2,$$

де $S_m^2 = \frac{1}{m-1}\sum_{i=1}^m (T_{r_i} - \overline{T})^2$, $T_{r_i} = \sum_{k \in s_{r_i}} y_k$, $i = \overline{1,m}$, $\overline{T} = \frac{1}{N}\sum_{i=1}^m T_{r_i}$.

4.2. Задачі

Задача 4.1. 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.

Задача 4.2. 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: $T_{s_1} = 8$ та $T_{s_2} = 12$.

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

Задача 4.3. 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. 4.1 наведено значення середньої успішності в школах кожної систематичної вибірки для кожного з трьох упорядкувань.

1) Підрахувати дисперсію оцінки Горвіца–Томпсона для середнього значення успішності при кожному із трьох упорядкувань.

Табл. 4.1: Середня успішність шкіл $\overline{y}$ для систематичних вибірок

Систематичні вибірки	уп. 1	уп. 2	уп. 3
1	8.83	8.50	8.65
2	8.93	9.06	8.67
3	8.53	9.50	9.36
4	11.82	9.73	8.77
5	11.05	10.14	9.17
6	9.23	9.67	9.50
7	9.59	9.04	9.12
8	8.43	8.53	9.44
9	7.04	8.49	9.63
10	8.42	9.21	9.56

2) Нехай дисперсія генеральної сукупності для середньої успішності $S^2 = 10.25$. Обчислити дисперсію оцінки Горвіца–Томпсона при ПВВбП з $n = 28$.

3) Обчислити коефіцієнт міри однорідності δ (див. формулу (4.2) на ст.73 в книзі [**?**]), для кожного з трьох упорядкувань. Яким буде мінімально можливе значення цього коефіцієнта?

4) При якому із цих упорядкувань є доцільним чи допустимим використання зміщеної оцінки для дисперсії (4.2)? А коли цією оцінкою не варто користуватись?

Задача 4.4. Пором, який перевозить автомобілі через затоку, стягує плату за один вантаж, а не з однієї особи. Поромна компанія хоче оцінити сумарну кількість людей, що вона перевозить в місяць. Вона вирішила провести обстеження в серпні за допомогою систематичного відбору. З минулого року компанія знає, що поромомом зазвичай користуються біля 400 автомобілів, і вони хочуть обстежити 80 автомобілів. Для полегшення оцінювання дисперсії систематичної вибірки дослідник обирає використання систематичної вибірки з 10 випадковими стартами. У цьому випадку утвориться 10 систематичних вибірок по 8 автомобілів у кожній.

Табл. 4.2: Номери машин та кількість людей в них для вибраних 10 систематичних вибірок

1 ел.	2 ел.	3 ел.	4 ел.	5 ел.	6 ел.	7 ел.	8 ел.
1(2)	51(1)	101(4)	151(8)	201(8)	251(3)	301(2)	351(3)
3(8)	53(5)	103(5)	153(6)	203(9)	253(9)	303(5)	353(6)
13(2)	63(2)	113(8)	163(5)	213(3)	263(6)	313(3)	363(6)
17(8)	67(1)	117(4)	167(4)	217(2)	267(6)	317(6)	367(6)
23(8)	73(1)	123(4)	173(4)	223(8)	273(4)	323(7)	373(6)
26(8)	76(4)	126(6)	176(7)	226(4)	276(9)	326(5)	376(8)
32(8)	82(1)	132(5)	182(5)	232(4)	282(3)	332(6)	382(2)
34(7)	84(1)	134(7)	184(6)	234(6)	284(4)	334(1)	384(9)
42(5)	92(7)	142(1)	192(8)	242(5)	292(7)	342(8)	392(3)
45(4)	95(7)	145(1)	195(2)	245(7)	295(8)	345(4)	395(3)

При реалізації такого систематичного відбору було відібрано 10 чисел від 1 до 50 випадковим чином із рівними ймовірностями і без повернення: 1, 3, 13, 17, 23, 26, 32, 34, 42, 45. У таблиці 5.4 подана інформація про порядковий номер автомобіля, який було обстежено та вказано кількість людей в ньому (у дужках).

Грунтуючись на даних, що наведені в таблиці 5.4, оцініть сумарну кількість осіб, що перевозить пором в місяць та побудуйте 95-% довірчий інтервал для сумарної кількості осіб. Квантилем якого розподілу тут більш доречно буде скористатись для побудови довірчого інтервалу (нормального чи розподілу Стьюдента)?

Задача 4.5. Показати, що при вибірковому дизайні СВ, коли $a = \frac{N}{n}$ – ціле число, дисперсія π-оцінки має вигляд

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

де $f = \frac{n}{N} = \frac{1}{a}$, δ - коефіцієнт міри однорідності δ, що визначений формулою (4.2) в книзі [**?**].

Розділ 5

Відбір з поверненням

5.1. Основний теоретичний матеріал

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 та всі N елементів беруть участь у відборі на кожному кроці. Такий відбір називають *простим випадковим відбором з поверненням* (англ. `simple random sampling with replacement`), скорочено ПВВзП.

При розгляді відбору з поверненням мають справу із впорядкованими вибірками. 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.*

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

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

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 $\sum\limits_{k \in U} p_k = 1$. Тоді процедура відбору проводиться так, що на першому кроці

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

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 $(k_1, k_2, ..., k_m)$

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

Оцінка Хансена–Гурвіца

Оцінкою Хансена–Гурвіца для сумарного T називається

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

Скорочення *pwr* походить від англійського виразу `p-expended with replacement`. Зауважимо, що сума в (5.1) береться по всіх елементах, що потрапили у впорядковану вибірку, незважаючи на повтори.

Оцінка Хансена–Гурвіца є незміщеною оцінкою сумарного T та має дисперсію

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

яку можна незміщено оцінити

$$\widehat{\mathcal{D}}(\widehat{t}_{pwr}) = \frac{1}{m} \widehat{V}_1, \text{ де } \quad \widehat{V}_1 = \frac{1}{m-1} \sum_{i=1}^{m} \left(\frac{y_{ki}}{p_{ki}} - \widehat{t}_{pwr} \right)^2. \tag{5.2}$$

Оцінка Хансена–Гурвіца при ПВВзП

При простому випадковому відборі з поверненням оцінка Хансена–Гурвіца для сумарного T має вигляд

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

де $\overline{y}_{os} = \frac{1}{m} \sum\limits_{i=1}^{m} y_{k_i}$ – середнє впорядкованої вибірки. Дисперсія та її оцінка при цьому мають вигляд:

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

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

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

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 вони були пропорційними до допоміжної змінної x, а саме $p_k = \frac{x_k}{\sum\limits_{i\in U} x_i}$, $k \in U$. Таким чином вибрані ймовірності називаються *ймовірностями, пропорційними до розміру* (англ. `probability proportional-to-size`), оскільки найчастіше характеристика x є деякою мірою розміру k-го елемента. В цьому випадку відбір називається *p-пропорційним до розміру* (англ. p-`proportional-to-size sampling`, або скорочено *pps*).

Для того, щоб отримати вибірку елементів, що потрапляють у неї із заданими ймовірностями p_k можна скористатись, наприклад, такими двома методами: *методом накопичених сум* та *методом 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 $T_{k-1} < \varepsilon T_N \leq T_k$, 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:

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

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.

5.2. Задачі

Задача 5.1. Розглянемо випадок простого випадкового відбору з поверненням з m відборами з генеральної сукупності розміру N. Довести, що $\mathrm{P}\{k \in os\} = \frac{m}{N} + O\left(\frac{m^2}{N^2}\right)$ коли N набуває великих значень.

Задача 5.2. 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: $\overline{y}_1 = \frac{1}{3}\sum_{k\in os} y_k$ – середнє з повторами; $\overline{y}_2 = \frac{1}{n_s}\sum_{k\in s} y_k$ – середнє по різних елементах із вибірки. Порахувати математичне сподівання та дисперсію цих оцінок. Який висновок можна зробити?

Задача 5.3. Яким повинен бути необхідний розмір впорядкованої вибірки m при ПВВзП, якщо обстежується генеральна сукупність, що складається із N елементів (N досить велике)? Нам потрібно оцінити пропорцію із точністю 1% та надійністю 95%.

Задача 5.4. В обстеженні кількості диких тварин досліджують територію площею 100 км^2, що поділена на смуги шириною 1 км,

але довжина в них різна. Вибірка 4-х смуг була зроблена за допомогою відбору з поверненням пропорційного до площі смуг. В кожній обраній смузі було підраховано кількість y тварин, що там проживають. Результати обстеження наступні:

Обрана смуга	Довжина смуги (км)	Кількість тварин
7	5	60
7	5	60
3	2	14
56	1	1

Порахуйте ймовірності p_k для обраних смуг та оцініть загальну кількість тварин, що проживають на обстежуваній території за допомогою оцінки Хансена-Гурвіца $\widehat{t}_{pwr}$ та оцінки Горвіца-Томпсона $\widehat{t}_{\pi}$. Оцініть дисперсію обох оцінок. Які переваги та недоліки кожної із цих оцінок?

Задача 5.5. Для генеральної сукупності U, що складається із 10 елементів, відома допоміжна змінна x: $x_1 = 10$, $x_2 = 9$, $x_3 = 4$, $x_4 = 7$, $x_5 = 7$, $x_6 = 8$, $x_7 = 8$, $x_8 = 2$, $x_9 = 2$, $x_{10} = 9$. Отримати вибірку із поверненням із 4 елементів, що потрапляють у неї із заданими ймовірностями $p_k = \frac{x_k}{\sum_{i \in U} x_i}$ двома методами: *методом накопичених сум* та *методом Лахірі*.

Розділ 6

Методи нерівноймовірнісного відбору без повернення

6.1. Основний теоретичний матеріал

До методів нерівноймовірнісного відбору, при яких елементи не мають жодного шансу потрапити у вибірку більше одного разу, належать відбір Пуассона та відбір, π-пропорційний до розміру.

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,

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

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.

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

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

Незміщенною оцінкою цієї дисперсії є статистика

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

Набір імовірностей $\pi_1, ..., \pi_N$ буде оптимальним, якщо відома деяка додаткова інформація (характеристика x), що досить добре корелює з y, то π_k вибираються 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 $\pi_k = 1$. Ймовірності включення, визначені формулою (6.1), є *ймовірностями, пропорційними до розміру* (англ. `probability proportional-to-size`).

Відбір, π-пропорційний до розміру

При відборі, π-пропорційному до розміру без повернення та з фіксованим розміром вибірки, ймовірності включення π_k вибираються пропорційно до деякої допоміжної змінної x, яка набуває додатних значень $x_1, x_2, ..., x_N$ та корелює з характеристикою y, що вивчається.

Реалізувати такий відбір набагато важче, ніж ПВВбП. Розглянемо дві схеми відбору, що дозволяють це зробити для розміру вибірки n: схему Брюера та схему Сантера.

Схема Брюера. [13] Ця схема розроблена для отримання вибірки розміру $n = 2$. Спочатку для всіх елементів k підраховуються величини

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

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

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

В схемі Сантера, яка широко використовується на практиці, 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 [22, 23]. 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 так: 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

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

6.2. Задачі

Задача 6.1. 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 $\widehat{t}_\pi$;

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

Задача 6.2. 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 $\widehat{\overline{y}}_\pi$: $cv(\widehat{\overline{y}}_\pi) = \frac{\sqrt{\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi)}}{\widehat{\overline{y}}_\pi}$.

Задача 6.3. 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 1.

Задача 6.4. 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}.

Задача 6.5. 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. Основний теоретичний матеріал

При *стратифікованому відборі (СТВ)* генеральна сукупність ділиться на підсукупності, що не перетинаються. Ці підсукупності називаються *стратами* (англ. `stratum`). А саме, нехай генеральна сукупність $U = \{1, \ldots, k, \ldots, N\}$ поділена на H страт: $U_1, \ldots, U_h, \ldots, U_H$, 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 $p_h(\cdot)$, $h = 1, \ldots, H$, 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

$$s = s_1 \cup s_2 \cup \ldots \cup s_H.$$

Внаслідок незалежності відборів зі страт вибірковий дизайн для стратифікованого відбору має вигляд:

$$p(s) = p_1(s_1)p_2(s_2)\ldots p_H(s_H).$$

Будемо вважати, що кількість елементів в h-й страті відома. Позначимо її через N_h, а через $W_h = \frac{N_h}{N}$ – вагу страти U_h.

При стратифікованому відборі π-оцінка сумарного значення T має вигляд

$$\widehat{t}_\pi = \sum_{h=1}^{H} \widehat{t}_{h\pi}, \tag{7.1}$$

де $\widehat{t}_{h\pi}$ – це незалежні π-оцінки сумарного значення в кожній із h-й страт. Дисперсія оцінки $\mathcal{D}_{\text{СТВ}}(\widehat{t}_\pi)$ та оцінка цієї дисперсії $\widehat{\mathcal{D}}_{\text{СТВ}}(\widehat{t}_\pi)$ при стратифікованому відборі знаходяться виходячи із незалежності відборів в кожній страті.

Стратифікований простий випадковий відбір

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).*

Нехай n_h – фіксований розмір простої випадкової вибірки з h-ї страти, $h = 1, \ldots, H$. При СТПВВ π-оцінка сумарного значення T має вигляд

$$\widehat{t}_\pi = \sum_{h=1}^{H} N_h \overline{y}_h, \tag{7.2}$$

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:

$$\mathcal{D}_{\text{СТПВВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_h^2, \tag{7.3}$$

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

$$\widehat{\mathcal{D}}_{\text{СТПВВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} \widehat{S}_h^2, \tag{7.4}$$

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.

Методи розміщень елементів вибірки по стратам при СТПВВ

Пропорційне розміщення визначається за формулою

$$n_h = \frac{n N_h}{N}, \quad h = 1, \ldots, H. \tag{7.5}$$

Оскільки ми вважаємо, що розміри страт N_h відомі заздалегідь, то таке розміщення завжди можна обчислити.

Розміщення Неймана визначається як

$$n_h = n \frac{N_h S_h}{\sum\limits_{l=1}^{H} N_l S_l}, \quad h = 1, \ldots, H. \tag{7.6}$$

Для знаходження значень n_h за цією формулою потрібно знати стандартні відхилення в стратах, тобто S_h, $h = 1, \ldots, H$.

Оптимальне розміщення. Припустимо, що загальні витрати на вибіркове обстеження можна зобразити у вигляді лінійної функції $C = c_0 + \sum_{h=1}^{H} n_h c_h$, де c_0 – деякі фіксовані витрати, а $c_h > 0$ – витрати на обстеження одного елемента h-ї страти. Мінімільну дисперсію для оцінки Горвіца-Томсона будемо мати при фіксованих витратах якщо

$$n_h = (C - c_0) \frac{N_h S_h / \sqrt{c_h}}{\sum_{l=1}^{H} N_l S_l \sqrt{c_l}}. \tag{7.7}$$

7.2. Задачі

Задача 7.1. Розглянемо генеральну сукупність, що складається із чотирьох елементів $U = \{1, 2, 3, 4\}$, для якої відомі значення характеристики y: $y_1 = y_2 = 0$, $y_3 = 1$, $y_4 = -1$. Досить природнім здається таке розбиття на страти: $U_1 = \{1, 2\}$ та $U_2 = \{3, 4\}$.

1. Порахувати дисперсію оцінки середнього для цієї сукупності при використанні ПВВбП із розміром вибірки $n = 2$.

2. Порахувати дисперсію оцінки середнього для цієї сукупності при використанні стратифікованого відбору із зазначеними вище стратами та обстеженні по одному елементу із кожної з них.

Порівняти та проаналізувати отримані результати.

Задача 7.2. Директор цирку має 100 слонів, які можна розділити на 2 категорії за статтю: слони та слонихи. Директору потрібно оцінити загальну вагу своїх слонів, оскільки їм потрібно переправитись через річку на паромі. Але це дуже трудоємкий процес, і директор це знає, оскільки того року він вже переважував всіх своїх слонів. Тогорічні результати подані в таблиці, де вага подана в тонах:

Страти	N_h	$\overline{Y}_h$	S_h^2
$h = 1$ (слони)	60	6	4
$h = 2$ (слонихи)	40	4	2.25

1. Порахувати дисперсію ваги всіх слонів у сукупності S^2.

2. Припустимо, що вага слонів не змінилась суттєво в кожній категорії за один рік. Якщо директор проведе відбір 10 слонів для зважування за допомогою ПВВбП, якою тоді буде дисперсія оцінки сумарної ваги?

3. Порахувати дисперсію оцінки сумарного, якщо директор проведе стратифікований відбір 10 слонів та використає *пропорційне розміщення* елементів по стратам.

4. Порахувати дисперсію оцінки сумарного, якщо директор проведе стратифікований відбір 10 слонів, але вже використає *розміщення Неймана* для розподілу елементів по стратам.

Проаналізувати виграш в точності для вказаних трьох відборів.

Задача 7.3. Проводиться обстеження підприємств однієї дуже великої країни. Однією із характеристик, що цікавить дослідників, є середній вік директорів підприємств. Сукупність підприємств поділена на три страти: малі, середні та великі підприємства. В кожній страті був проведений ПВВбП відбір. Інформація про вагу страти $W_h = N_h/N$, вибіркове середнє значення віку директорів $\overline{y}_h$, дисперсія віку S_h^2, кількість підприємств n_h, що були обстежені та вартість обстеження c_h (в умовних одиницях) в кожній страті подана в таблиці нижче.

Страти	W_h	$\overline{y}_h$	S_h^2	n_h	c_h
малі підприємства	60%	30	16	40	1
середні підприємства	30%	45	10	20	1
великі підприємства	10%	55	20	40	4

1. Оцініть середній вік директорів для всієї сукупності підприємств за допомогою π-оцінки. Чи співпадає вона із середнім вибірковим?

2. Обчислити дисперсію π-оцінки середнього віку, якщо вважати, що частка відобору в кожній страті n_h/N_h настільки мала, що нею можна знехтувати.

3. Порахувати яким було б пропорційне розміщення 100 елементів в цьому обстеженні. Яку б дисперсію в цьому випадку ми б отримали?

4. За умови однакових фіксованих сумарних витрат на обстеження, що були зроблені, порахувати яким би було оптимальне розміщення та визначити на скільки суттєвими були б зміни в точності в порівнянні із початковим розміщенням.

Задача 7.4. У деякому вибірковому обстеженні використовується СТПВВ. Припустимо, що потрібно оцінити середнє значення $\overline{Y}$ досліджуваної характеристики генеральної сукупності з такою точністю, щоб умова

$$\left| \sum_{h=1}^{H} W_h \overline{y}_h - \overline{Y} \right| \leq a$$

виконувалася при заданому значенні сталої a щонайменше з імовірністю $1 - \alpha$. Показати, що розмір вибірки, необхідний для виконання цієї умови, задовольняє нерівність

$$n \geq \frac{\left(\frac{z}{a}\right)^2 \sum\limits_{h=1}^{H} \frac{W_h^2 S_h^2}{w_h}}{1 + \frac{1}{N}\left(\frac{z}{a}\right)^2 \sum\limits_{h=1}^{H} W_h S_h^2},$$

де $w_h = n_h/n$, $z = z_{1-\alpha/2}$.

Задача 7.5. Нехай в умові попередньої задачі $N = 1000$, $H = 2$, $W_1 = 1 - W_2 = 0.8$, $S_1^2 = 4$, $S_2^2 = 16$, $z = 1.96$, $a = 0.5$. Дослідити, як змінюється необхідний розмір вибірки n як функція від $w_1 = 1 - w_2$, якщо w_1 змінюється від 0 до 1.

Розділ 8

Кластерний, двостадійний та багатостадійний відбори

8.1. Основний теоретичний матеріал

Кластерний відбір (КВ) (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*.

У випадку *одностадійного кластерного відбору* (ОКВ) спочатку відбирають кластери, а потім обстежують всі елементи відібраних кластерів.

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 та такий двостадійний відбір називають *двостадійним кластерним відбором* (ДКВ) (англ. `two-stage cluster sampling`). Якщо всі вторинні вибіркові одиниці – це окремі елементи, то такий двостадійний відбір ще називають *двостадійним відбором елементів* (ДВЕ) (англ. `two-stage element sampling`).

Якщо процедура відбору складається з трьох або більше стадій, то такий відбір називається *багатостадійним*.

Одностадійний кластерний відбір.

В цьому випадку скінченна генеральна сукупність $U = \{1, 2, ..., N\}$ ділиться на N_{I} підмножин (кластерів):

$$U_1, U_2, ..., U_{N_{\text{I}}}.$$

Відбір реалізується таким чином:

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 $s = \bigcup\limits_{i \in s_{\mathrm{I}}} U_i$. Розмір вибірки s дорівнює $n_s = \sum\limits_{s_{\mathrm{I}}} N_i$.

Зауважимо, що навіть якщо $p_{\mathrm{I}}(\cdot)$ є вибірковим дизайном із фіксованим розміром вибірки кластерів, то розмір вибірки елементів не обов'язково буде фіксованим.

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

$$\pi_{\mathrm{I}ij} = \sum_{s_{\mathrm{I}} \ni i \& j} p_{\mathrm{I}}(s_{\mathrm{I}}), \quad i, j = 1, \ldots, N_{\mathrm{I}}.$$

Оскільки вибірка s містить всі елементи відібраних кластерів, то для довільного елемента k з кластера U_i будемо мати

$$\pi_k = \mathrm{P}(k \in s) = \mathrm{P}(i \in s_{\mathrm{I}}) = \pi_{\mathrm{I}i}. \tag{8.1}$$

Ймовірність включення другого порядку дорівнює $\pi_{kl} = \mathrm{P}(k \& l \in s) = \pi_{\mathrm{I}i}$, якщо обидва елементи k та l належать одному кластеру U_i та $\pi_{kl} = \mathrm{P}(k \& l \in s) = \mathrm{P}(i \& j \in s_{\mathrm{I}}) = \pi_{\mathrm{I}ij}$, якщо елементи k та l належать різним кластерам: $k \in U_i$, $l \in U_j$, $i, j = 1, \ldots, N_{\mathrm{I}}$.

Оцінка Горвіца-Томсона при ОКВ.

Нехай $T_i = \sum\limits_{k \in U_i} y_k$ – сумарне значення досліджуваної характеристики в i-му кластері. При ОКВ π-оцінка сумарного значення має вигляд

$$\widehat{t}_\pi = \sum_{i \in s_\mathrm{I}} \frac{T_i}{\pi_{\mathrm{I}i}}. \tag{8.2}$$

Дисперсія цієї оцінки дорівнює

$$\mathcal{D}(\widehat{t}_\pi) = \sum_{i \in U_\mathrm{I}} \sum_{j \in U_\mathrm{I}} (\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}) \frac{T_i}{\pi_{\mathrm{I}i}} \frac{T_j}{\pi_{\mathrm{I}j}}. \tag{8.3}$$

Незміщена оцінка дисперсії обчислюється за формулою

$$\widehat{\mathcal{D}}(\widehat{t}_\pi) = \sum_{i \in s_\mathrm{I}} \sum_{j \in s_\mathrm{I}} \frac{\pi_{\mathrm{I}ij} - \pi_{\mathrm{I}i}\pi_{\mathrm{I}j}}{\pi_{\mathrm{I}ij}} \frac{T_i}{\pi_{\mathrm{I}i}} \frac{T_j}{\pi_{\mathrm{I}j}}. \tag{8.4}$$

Простий випадковий одностадійний кластерний відбір 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 без повернення, після чого обстежуються всі елементи відібраних кластерів.

Тоді

$$\widehat{t}_\pi = N_\mathrm{I} \overline{T}_{s_\mathrm{I}}, \tag{8.5}$$

де $\overline{T}_{s_\mathrm{I}} = \frac{1}{n_\mathrm{I}} \sum\limits_{i \in s_\mathrm{I}} T_i$ – середнє арифметичне сумарних значень досліджуваної характеристики, обчислене за тими кластерами, які потрапили до вибірки s_I.

Дисперсія оцінки сумарного значення обчислюється за формулою

$$\mathcal{D}_{\text{ПВОКВ}}(\widehat{t}_\pi) = N_\mathrm{I}^2 \frac{1 - f_\mathrm{I}}{n_\mathrm{I}} S^2_{TU_\mathrm{I}}, \tag{8.6}$$

де $f_\mathrm{I} = \frac{n_\mathrm{I}}{N_\mathrm{I}}$ – частка відбору кластерів,

$S^2_{TU_\mathrm{I}} = \frac{1}{N_\mathrm{I}-1} \sum\limits_{i \in U_\mathrm{I}} \left(T_i - \overline{T}_{U_\mathrm{I}}\right)^2$ – дисперсія сумарних значень у кластерах,

$\overline{T}_{U_\mathrm{I}} = \frac{1}{N_\mathrm{I}} \sum\limits_{i \in U_\mathrm{I}} T_i$ – середнє арифметичне сумарних значень досліджуваної характеристики, обчислене за всіма кластерами генеральної сукупності.

Незміщена оцінка дисперсії $\mathcal{D}_{\text{ПВОКВ}}(\widehat{t_\pi})$ дорівнює

$$\widehat{\mathcal{D}}_{\text{ПВОКВ}}(\widehat{t_\pi}) = N_\mathrm{I}^2 \frac{1 - f_\mathrm{I}}{n_\mathrm{I}} S^2_{T s_\mathrm{I}}, \tag{8.7}$$

де $S^2_{T s_\mathrm{I}} = \frac{1}{n_\mathrm{I} - 1} \sum\limits_{i \in s_\mathrm{I}} \left(T_i - \overline{T}_{s_\mathrm{I}}\right)^2$.

При двостадійному відборі загальна процедура відбору полягає в наступному.

Перша стадія. Відбирається вибірка s_I первинних вибіркових одиниць із множини U_I ($s_\mathrm{I} \subset U_\mathrm{I}$) згідно з вибірковим дизайном $p_\mathrm{I}(\cdot)$.

Друга стадія. Для кожного $i \in s_\mathrm{I}$ відбирається вибірка елементів, яку ми позначимо s_i, з первинної вибіркової одиниці U_i ($s_i \subset U_i$) згідно з вибірковим дизайном $p_i(\cdot|s_\mathrm{I})$. Результуюча вибірка елементів, яку позначимо як s, є об'єднанням вибірок з відібраних на першій стадії первинних вибіркових одиниць: $s = \bigcup\limits_{i \in s_\mathrm{I}} s_i$.

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 зазвичай обмежуються розглядом двостадійних відобрів, коли виконуються умови інваріантності, тобто відбір в кожній ПВО не залежить від інших ПВО, та незалежності, коли відбір на першій та другій стадіях не впливають один на одного.

При *двостадійному кластерному відборі* вторинними вибірковими одиницями є елементи.

Для вибіркового дизайну $p_\mathrm{I}(\cdot)$ першої стадії відбору ймовірності включення ПВО у вибірку позначаються $\pi_{\mathrm{I}i}$ та $\pi_{\mathrm{I}ij}$. 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:

$$\pi_k = \pi_{\mathrm{I}i}\pi_{k|i}, \quad \text{де} \quad i \in U_{\mathrm{I}} : U_i \ni k, \tag{8.8}$$

а ймовірність того, що k-й та l-й елементи генеральної сукупності потраплять до вибірки, дорівнює

$$\pi_{kl} = \begin{cases} \pi_{\mathrm{I}i}\pi_{k|i}, & \text{якщо} \quad k = l \in U_i; \\ \pi_{\mathrm{I}i}\pi_{kl|i}, & \text{якщо} \quad k \& l \in U_i, k \neq l; \\ \pi_{\mathrm{I}ij}\pi_{k|j}\pi_{l|j}, & \text{якщо} \quad k \in U_i, l \in U_j, (i \neq j). \end{cases} \tag{8.9}$$

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:

$$\widehat{t}_{i\pi} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}}. \tag{8.10}$$

Дисперсія цієї оцінки обчислюється за формулою

$$\mathcal{D}_i = \sum_{k \in U_i} \sum_{l \in U_i} \left(\pi_{kl|i} - \pi_{k|i}\pi_{l|i}\right) \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}, \tag{8.11}$$

а її незміщена оцінка дорівнює

$$\widehat{\mathcal{D}}_i = \sum_{k \in s_i} \sum_{l \in s_i} \frac{\left(\pi_{kl|i} - \pi_{k|i}\pi_{l|i}\right)}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}. \tag{8.12}$$

Оцінка Горвіца-Томсона при двостадійному відборі елементів сумарного значення $T = \sum\limits_{k \in U} y_k$ має вигляд

$$\widehat{t}_{\pi} = \sum_{i \in s_{\mathrm{I}}} \frac{\widehat{t}_{i\pi}}{\pi_{\mathrm{I}i}}. \tag{8.13}$$

Дисперсія оцінки $\widehat{t}_{\pi}$ має дві складові:

$$\mathcal{D}\left(\widehat{t}_{\pi}\right) = \mathcal{D}_{\text{ПВО}} + \mathcal{D}_{\text{ВВО}}, \tag{8.14}$$

де

$$\mathcal{D}_{\text{ПВО}} = \sum_{i \in U_{\text{I}}} \sum_{j \in U_{\text{I}}} (\pi_{\text{I}ij} - \pi_{\text{I}i}\pi_{\text{I}j}) \frac{T_i}{\pi_{\text{I}i}} \frac{T_j}{\pi_{\text{I}j}}, \tag{8.15}$$

$$\mathcal{D}_{\text{ВВО}} = \sum_{i \in U_{\text{I}}} \frac{\mathcal{D}_i}{\pi_{\text{I}i}}. \tag{8.16}$$

Незміщеною оцінкою дисперсії $\mathcal{D}_{\text{ПВО}}$ є статистика

$$\begin{aligned} \widehat{\mathcal{D}}_{\text{ПВО}} &= \sum_{i \in s_{\text{I}}} \sum_{j \in s_{\text{I}}} \frac{(\pi_{\text{I}ij} - \pi_{\text{I}i}\pi_{\text{I}j})}{\pi_{\text{I}ij}} \frac{\widehat{t}_{i\pi}}{\pi_{\text{I}i}} \frac{\widehat{t}_{j\pi}}{\pi_{\text{I}j}} - \\ &- \sum_{i \in s_{\text{I}}} \frac{1}{\pi_{\text{I}i}} \left(\frac{1}{\pi_{\text{I}i}} - 1 \right) \widehat{\mathcal{D}}_i \,, \end{aligned} \tag{8.17}$$

а незміщена оцінка дисперсії $\mathcal{D}_{\text{ВВО}}$ дорівнює

$$\widehat{\mathcal{D}}_{\text{ВВО}} = \sum_{i \in s_{\text{I}}} \frac{\widehat{\mathcal{D}}_i}{(\pi_{\text{I}i})^2}. \tag{8.18}$$

Отже, незміщена оцінка дисперсії $\mathcal{D}\left(\widehat{t}_\pi\right)$ має вигляд

$$\begin{aligned} \widehat{\mathcal{D}}\left(\widehat{t}_\pi\right) &= \widehat{\mathcal{D}}_{\text{ПВО}} + \widehat{\mathcal{D}}_{\text{ВВО}} = \\ &= \sum_{i \in s_{\text{I}}} \sum_{j \in s_{\text{I}}} \frac{(\pi_{\text{I}ij} - \pi_{\text{I}i}\pi_{\text{I}j})}{\pi_{\text{I}ij}} \frac{\widehat{t}_{i\pi}}{\pi_{\text{I}i}} \frac{\widehat{t}_{j\pi}}{\pi_{\text{I}j}} + \sum_{i \in s_{\text{I}}} \frac{\widehat{\mathcal{D}}_i}{(\pi_{\text{I}i})}. \end{aligned} \tag{8.19}$$

Простий випадковий відбір на обох стадіях двостадійного відбору елементів. 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 $i \in s_{\text{I}}$ 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

$$\widehat{t}_\pi = \frac{N_{\text{I}}}{n_{\text{I}}} \sum_{i \in s_{\text{I}}} N_i \overline{y}_i = \frac{N_{\text{I}}}{n_{\text{I}}} \sum_{i \in s_{\text{I}}} \widehat{t}_{i\pi}, \tag{8.20}$$

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

$$\mathcal{D}_{\text{ПВДВЕ}} \left(\widehat{t}_\pi \right) = {N_{\text{I}}}^2 \frac{1 - f_{\text{I}}}{n_{\text{I}}} S_T^2 + \frac{N_{\text{I}}}{n_{\text{I}}} \sum_{i \in U_{\text{I}}} N_i^2 \frac{1 - f_i}{n_i} S_i^2, \qquad (8.21)$$

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

$$\widehat{\mathcal{D}}_{\text{ПВДВЕ}} \left(\widehat{t}_\pi \right) = {N_{\text{I}}}^2 \frac{1 - f_{\text{I}}}{n_{\text{I}}} \widehat{S}_T^2 + \frac{N_{\text{I}}}{n_{\text{I}}} \sum_{i \in s_{\text{I}}} N_i^2 \frac{1 - f_i}{n_i} \widehat{S}_i^2, \qquad (8.22)$$

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,

$\widehat{S}_i^2 = \frac{1}{n_i - 1} \sum\limits_{k \in s_i} \left(y_k - \overline{y}_i \right)^2$ – вибіркова дисперсія досліджуваної характеристики в i-й ПВО.

8.2. Задачі

Задача 8.1. 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.

3) Оцінити середній дохід домогосподарства району та дисперсію отриманої оцінки.

Задача 8.2. 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 частку 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.

Задача 8.3. 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 $n_{\mathrm{I}} = 5$ з $N_{\mathrm{I}} = 50$ 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, $i \in s_{\mathrm{I}}$. 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.

Задача 8.4. У невеликому коледжі мистецтв є 36 відділень. Потрібно оцінити середню суму грошей, яку студенти витратили на робочі матеріали минулого семестру. Оскільки розмір кожного відділення дуже різний, вирішено було проводити обстеження в

два етапи. На першому етапі відбираються відділення із ймовірностями пропорційно їх розміру із поверненням. На другому етапі відбирається вибірка студентів ПВВбП в кожному відділенні, обраному на першому етапі. Результати наведені в таблиці нижче.

i	N_i	n_i	Витрати
1	10	4	326, 400, 423, 443
2	20	8	278, 312, 450, 350, 227, 438, 512, 403
3	30	12	512, 256, 332, 402, 512, 309, 411, 610, 422, 630, 550, 470
4	15	6	426, 312, 512, 440, 342, 533

1. Випишіть який вигляд буде мати оцінка Хансена-Гурвіца для середнього рівня витрат студентів при такому вибірковому дизайні.

2. Як можна оцінити дисперсію такої оцінки?

3. Побудуйте 95% довірчий інтервал для шуканого середнього рівня витрат студентів.

Розділ 9

Оцінювання функцій від сумарних значень характеристик генеральної сукупності

9.1. Основний теоретичний матеріал

Оцінювання вектора сумарних значень

Нехай досліджуються q характеристик генеральної сукупності, які ми будемо позначати через $y_1, \ldots, y_j, \ldots, y_q$. Значення цих характеристик для N елементів генеральної сукупності будемо позначати через $y_{j1}, \ldots, y_{jk}, \ldots, y_{jN}$, $j = 1, \ldots, q$. Потрібно оцінити q компонент вектора невідомих сумарних значень цих характеристик:

$$\mathbf{T} = (T_1, \ldots, T_j, \ldots, T_q)', \quad \text{де} \quad T_j = \sum_{k \in U} y_{jk}.$$

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

$$\mathbf{y}_k = (y_{1k}, \ldots, y_{jk}, \ldots, y_{qk}).$$

Нехай сумарне значення кожної досліджуваної характеристики оцінюється за допомогою оцінки Горвіца–Томпсона (π-оцінки). Тоді вектор оцінок сумарних значень буде мати вигляд

$$\widehat{\mathbf{t}}_\pi = (\widehat{t}_{1\pi}, \ldots, \widehat{t}_{j\pi}, \ldots, \widehat{t}_{q\pi})', \quad \text{де} \quad \widehat{t}_{j\pi} = \sum_{k \in s} \frac{y_{jk}}{\pi_k}.$$

Вектор $\widehat{\mathbf{t}}_\pi$ є незміщеною оцінкою вектора $\mathbf{T}$ з коваріаційною матрицею $\mathbf{V}(\widehat{\mathbf{t}}_\pi)$, елементи якої задають коваріацію оцінок $\widehat{t}_{i\pi}$ та $\widehat{t}_{j\pi}$ $(i, j = 1, \ldots, q)$:

$$\mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) := \operatorname{cov}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_{ik}}{\pi_k} \frac{y_{jl}}{\pi_l}. \tag{9.1}$$

Незміщеною оцінкою матриці $\mathbf{V}(\widehat{\mathbf{t}}_\pi)$ є матриця $\widehat{\mathbf{V}}(\widehat{\mathbf{t}}_\pi)$ елементами якої є оцінки коваріацій

$$\widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_{ik}}{\pi_k} \frac{y_{jl}}{\pi_l}, \; i, j = \overline{1, q}. \tag{9.2}$$

Оцінювання функцій від сумарних значень кількох змінних

Припустимо, що потрібно оцінити параметр θ генеральної сукупності U, який можна подати у вигляді функції від q сумарних значень $T_1, \ldots, T_q$:

$$\theta = f(T_1, \ldots, T_q), \quad \text{де} \quad T_j = \sum_{k \in U} y_{jk}, \quad j = 1, \ldots, q.$$

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

$$\hat{\theta} = f(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}). \tag{9.3}$$

Якщо *функція f - лінійна*, тобто $\theta = a_0 + \sum\limits_{j=1}^{q} a_j T_j$, то оцінка $\hat{\theta} = a_0 + \sum\limits_{j=1}^{q} a_j \widehat{t}_{j\pi}$ є незміщеною оцінкою параметра θ, а дисперсія цієї оцінки

$$\mathcal{D}(\hat{\theta}) = \mathcal{D}\left(\sum_{j=1}^{q} a_j \widehat{t}_{j\pi}\right) = \sum_{i=1}^{q}\sum_{j=1}^{q} a_i a_j \mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}), \tag{9.4}$$

де коваріація $\mathbf{C}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi})$ визначається за формулою (9.1).

Оцінкою дисперсії (9.4) є статистика

$$\widehat{\mathcal{D}}(\hat{\theta}) = \sum_{i=1}^{q}\sum_{j=1}^{q} a_i a_j \widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi}), \tag{9.5}$$

де оцінка $\widehat{\mathbf{C}}(\widehat{t}_{i\pi}, \widehat{t}_{j\pi})$ визначається за формулою (9.2).

Метод лінеаризації Тейлора для оцінювання дисперсії використовується коли $\theta = f(T_1, \ldots, T_q)$ є *нелінійною функцією* від q сумарних значень $T_1, \ldots, T_q$ і неможливо отримати точні значення зміщення та дисперсії оцінки $\hat{\theta} = f(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi})$. Цей метод полягає у наближенні нелінійної оцінки $\hat{\theta}$ псевдооцінкою $\hat{\theta}_0$, яка є лінійною функцією від змінних $\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}$ та отримується із розкладу Тейлора функції f в околі точки $\mathbf{T}$. Якщо це наближення

досить точне, то замість дисперсії $\mathcal{D}(\hat{\theta})$ можна використовувати дисперсію $\mathcal{D}(\hat{\theta}_0)$, яка обчислюється порівняно легко.

«Приблизно незміщеною» оцінкою параметра $\theta = f(T_1, \ldots, T_q)$, де $T_1 = \sum\limits_{k \in U} y_{1k}, \ldots, T_q = \sum\limits_{k \in U} y_{qk}$, 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, 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 $u_k = \sum\limits_{j=1}^{q} a_j y_{jk}$, а коефіцієнти $a_j = \left. \frac{\partial f}{\partial \widehat{t}_{j\pi}} \right|_{(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}) = (T_1, \ldots, T_q)}$.

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 $\hat{u}_k = \sum\limits_{j=1}^{q} \hat{a}_j y_{jk}$, а коефіцієнти $\hat{a}_j$ отримані шляхом заміни невідомих сумарних значень відповідними π-оцінками.

У цьому твердженні ми використали вираз «*приблизно незміщена оцінка*», який означає, що зміщення запропонованої оцінки є настільки малим порівняно з її дисперсією, що ним можна знехтувати.

Оцінювання відношення – це задача, коли потрібно оцінити параметр, що є відношенням невідомих сумарних значень двох досліджуваних характеристик генеральної сукупності (не плутати із пропорцією P_d!):

$$R = \frac{T_y}{T_z} = \frac{\sum\limits_{k \in U} y_k}{\sum\limits_{k \in U} z_k}. \tag{9.6}$$

Наприклад, якщо U – це генеральна сукупність домогосподарств, y_k – сумарний дохід k-ого домогосподарства, а z_k – кількість осіб

в k-ому домогосподарстві, то R – це середній дохід у розрахунку на одну особу.

В якості оцінки відношення R розглядають $\widehat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}$, що є нелінійною функцією від π-оцінок. Ця оцінка є зміщеною, але для великих сукупностей та вибірок, зміщення буде незначним і ним зазвичай нехтують.

Результатом застосування *методу лінеаризації Тейлора* є лінійне наближення статистики

$$\widehat{R} \approx \widehat{R}_0 = R + \frac{1}{T_z} \sum_{k \in s} \frac{(y_k - Rz_k)}{\pi_k} = R + \frac{1}{T_z}(\hat{t}_{y\pi} - R\hat{t}_{z\pi}).$$

Оцінка $\widehat{R}$ є приблизно незміщеною оцінкою параметра R, а наближене значення дисперсії цієї оцінки:

$$\mathcal{D}(\widehat{R}) \approx \mathcal{D}(\widehat{R}_0) = \frac{1}{T_z^2} \left(\mathcal{D}(\hat{t}_{y\pi}) + R^2 \mathcal{D}(\hat{t}_{z\pi}) - 2R\mathbf{C}(\hat{t}_{y\pi}, \hat{t}_{z\pi}) \right). \quad (9.7)$$

В якості оцінки дисперсії оцінки $\widehat{R}$ розглядають статистику:

$$\widehat{\mathcal{D}}(\widehat{R}) = \frac{1}{\hat{t}_{z\pi}^2} \left(\widehat{\mathcal{D}}(\hat{t}_{y\pi}) + \widehat{R}^2 \widehat{\mathcal{D}}(\hat{t}_{z\pi}) - 2\widehat{R}\widehat{\mathbf{C}}(\hat{t}_{y\pi}, \hat{t}_{z\pi}) \right). \quad (9.8)$$

У випадку *простого випадкового відбору без повернення* з розміром вибірки $n = fN$: $\hat{t}_{y\pi} = N\overline{y}$, $\hat{t}_{z\pi} = N\overline{z}$ та $\widehat{R} = \frac{\overline{y}}{\overline{z}}$, де $\overline{y} = \frac{1}{n}\sum_{k \in s} y_k$, $\overline{z} = \frac{1}{n}\sum_{k \in s} z_k$ – вибіркові середні характеристик y та z відповідно. Тоді, вираз для наближеного обчислення дисперсії оцінки $\widehat{R}$ має вигляд:

$$\widetilde{\mathcal{D}}(\widehat{R}) = \frac{1}{(\overline{Z})^2} \frac{1-f}{n} \frac{1}{N-1} \sum_{k \in U} (y_k - Rz_k)^2 =$$

$$= \frac{1}{(\overline{Z})^2} \frac{1-f}{n} \left(S_y^2 + R^2 S_z^2 - 2RS_{yz} \right),$$

де S_{yz} – коваріація змінних y та z, обчислена за всією генеральною сукупністю. Оцінити дисперсію оцінки $\widehat{R}$ можна як:

$$\widehat{\mathcal{D}}(\widehat{R}) = \frac{1}{(\overline{z})^2} \frac{1-f}{n} \frac{1}{n-1} \sum_{k \in s} (y_k - \widehat{R}z_k)^2 = \quad (9.9)$$

$$= \frac{1}{(\overline{z})^2} \frac{1-f}{n} \left(\widehat{S}_y^2 + \widehat{R}^2 \widehat{S}_z^2 - 2\widehat{R}\widehat{S}_{yz} \right),$$

де

$$\widehat{S}_y^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \overline{y})^2, \quad \widehat{S}_z^2 = \frac{1}{n-1} \sum_{k \in s} (z_k - \overline{z})^2,$$

$$\widehat{S}_{yz} = \frac{1}{n-1} \sum_{k \in s} (y_k - \overline{y})(z_k - \overline{z}).$$

Метод Джек-найф (від англ. `Jack-knife` – складний ніж). Нехай s – це вибірка розміру n, що була отримана за допомогою деякого вибіркового дизайну $p(s)$, $\widehat{\theta}$ – оцінка параметра θ. Розглянемо $\widehat{\theta}_{(j)}$ – оцінку того ж типу, що і $\widehat{\theta}$, але пораховану по всіх елементах, крім j-того: по елементам із $s_{(j)} = s \setminus \{j\}$.

Розглянемо JK-оцінку для параметра θ виду

$$\widehat{\theta}_{JK} = \frac{1}{n} \sum_{j=1}^{n} \widehat{\theta}_{(j)}.$$

в якості оцінки для дисперсії $\mathcal{D}(\widehat{\theta})$ можна розглянути дві статистики:

$$\widehat{\mathcal{D}}_{JK1}(\widehat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} (\widehat{\theta}_{(j)} - \widehat{\theta}_{JK})^2$$

$$\widehat{\mathcal{D}}_{JK2}(\widehat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} (\widehat{\theta}_{(j)} - \widehat{\theta})^2.$$

При цьому $\widehat{\mathcal{D}}_{JK2} \geq \widehat{\mathcal{D}}_{JK1}$.

При ПВВбП розміру $n = fN$ параметр відношення R можна оцінити за допомогою оцінки $\widehat{R} = \frac{\sum_{k \in s} y_k}{\sum_{k \in s} z_k}$. Для JK-оцінки нам потрібно порахувати оцінки відношення по вибіркам, що не містять один елемент j, $j = \overline{1,n}$:

$$\widehat{R}_{(j)} = \frac{\sum_{k \neq j, k \in s} y_k}{\sum_{k \neq j, k \in s} z_k}.$$

Бутстреп метод (від англ. `bootstrap`, скорочено BS). Припустимо, що ймовірнісна вибірка s розміру n була отримана за допомогою деякого вибіркового дизайну $p(s)$ (дизайн без повернення) із генеральної сукупності U розміру N. Нехай $\widehat{\theta}$ – оцінка параметра θ. Потрібно оцінити дисперсію $\mathcal{D}(\widehat{\theta})$.

Застосування Бутстреп методу можна поділити на 3 етапи:

- Використовуючи дані y_k, отримані із вибірки s, ми конструюємо штучну генеральну сукупність U^* розміру N (якщо це не можливо, або хоча б приблизно такого ж розміру), що імітує справжню генеральну сукупність. Для цього використаємо ваги наших елементів у вибірці: якщо ймовірність включення для елемента k дорівнює π_k, тоді його ваги $w_k = \frac{1}{\pi_k}$. І таким чином елемент k буде представлений в штучній генеральній сукупності w_k разів.

- Робимо відбір серії K (K – має порядок декілька десятків тисяч) вибірок із U^* використовуючи початковий вибірковий дизайн $p(s)$ та підраховуємо оцінки $\widehat{\theta}^*_r$, $r = \overline{1, K}$, так само як підраховували $\widehat{\theta}$.

- Отриманий розподіл значень $\widehat{\theta}^*_1$, $\widehat{\theta}^*_2$, ... $\widehat{\theta}^*_R$ розглядається як наближення вибіркового розподілу $\widehat{\theta}$ та для оцінювання $\mathcal{D}(\widehat{\theta})$ використовується функція

$$\widehat{\mathcal{D}}_{BS}(\widehat{\theta}) = \frac{1}{K-1}\sum_{r=1}^{K}(\widehat{\theta}^*_r - \widehat{\theta}^*)^2, \quad \text{де} \quad \widehat{\theta}^* = \frac{1}{K}\sum_{r=1}^{K}\widehat{\theta}^*_r.$$

Якщо в задачі головним є побудова довірчого інтервала для $\widehat{\theta}$, тоді оцінювати дисперсію не обов'язково, а нижня та верхня межі довірчого інтервала беруться із розподілу $\widehat{\theta}^*_1$, $\widehat{\theta}^*_2$, ... $\widehat{\theta}^*_K$.

9.2. Задачі

Задача 9.1. Проводиться вибіркове обстеження, метою якого є оцінювання різниці між сумарними значеннями двох характеристик генеральної сукупності y та z. Розмір генеральної сукупності

$N = 2400$. За допомогою простого випадкового відбору без повернення отримано вибірку розміру $n = 40$ та підраховано:

$$\sum_{k\in s} y_k = 866;\ \widehat{S}_y^2 = 42.438;\ \sum_{k\in s} z_k = 383;\ \widehat{S}_z^2 = 29.430;\ \widehat{S}_{yz} = 1.309.$$

Оцінити різницю сумарних значень y та z: $D = T_y - T_z$ та побудувати 95% довірчий інтервал для цього параметра.

Задача 9.2. Спеціальний моніторинговий комітет планує досліджувати динаміку змін цін на пальне. Для опрацювання загальних підходів розглядаються (тренуються на) 10 заправочних станцій які в цьому випадку відіграють роль генеральної сукупності. Впродовж двох місяців (квітень та травень) відслідковувались ціни на найбільш розповсюджений тип пального на цих 10 станціях. Результати спостережень подані у таблиці

№ зп.станції	1	2	3	4	5
Квітень	25.82	25.33	25.76	25.98	26.20
Травень	25.89	25.34	25.92	26.05	26.20
№ зп.станції	6	7	8	9	10
Квітень	25.89	25.68	25.55	25.69	25.81
Травень	26.00	25.79	25.63	25.78	25.84

Якщо проводити вибіркове обстеження до цих 10 заправочних станцій ПВВбП із вибіркою розміру n та досліджувати ціни у травні (y_k) та у квітні (x_k), то можна розглянути два сценарії:

Сц. 1 Можемо отримувати **незалежні** вибірки розміру n: в квітні та в травні (окремо вибирається n станцій із 10 в квітні, і незалежно від того, які станції обстежуються в квітні, в травні робиться нова вибірка розміру n).

Сц. 2 Можемо в квітні отримати вибірку розміру n, та в травні обстежувати ті ж самі заправочні станції, що були обстежені в квітні. Чи будуть в цьому випадку вибірки **незалежні**?

Дайте відповідь на наступні питання:

1. Ми хочемо оцінити параметр $D = \bar{Y} - \bar{X}$. Знайдіть дисперсії оцінки різниці $\widehat{d} = \bar{y} - \bar{x}$ для першого та другого сценаріїв. Порівняйте отримані результати. Який сценарій буде більш ефективним?

2. Ми хочемо об'єднати обидва цих часових періоди в один комбінований часовий інтервал квітень-травень та оцінити середню ціну пального за цей об'єднаний період $\overline{XY} = \frac{1}{N+N}\left(\sum_{k\in U} x_k + \sum_{k\in U} y_k\right)$. Який в цьому випадку сценарій буде більш ефективним? Порахуйте дисперсії для цього випадку та порівняйте їх.

3. Припустимо, що ми в об'єднаному періоді квітень-травень будемо обирати вибірку розміру $2n$ не вимагаючи, щоб рівно n станцій було вибрано кожного місяця і будемо оцінювати параметр середнє. Тобто ми станції об'єднуємо і вважаємо для кожного місяця їх різними та вибираємо одночасно із 20 станцій $2n$ ПВВбП. Також будемо вважати, що дані отримані виходячи із сценарію 1. Що ви можете сказати щодо точності оцінювання середнього в цьому випадку в порівнянні із п. 2?

Задача 9.3. В деякому місті Статислав потрібно оцінити відсоток жінок серед осіб віком 60+. Для цього серед 1024 домогосподарств (дг), що знаходяться в цьому місті було вибрано ПВВбП та обстежено 16 дг. Результати обстеження наступні: лише в 3 дг проживали особи віком 60+, при чому у 2-х дг проживало по двоє таких осіб, одна з яких була жінкою, а у одному дг проживала одна жінка віком 60+. Всі інші дг не мали в своєму складі осіб віком 60+.

Побудувати 95% довірчі інтервали для шуканого параметра R використовуючи для оцінювання дисперсії: а) метод ліанеризації Тейлора; б) метод Джек-найф; в) метод Бут-стреп.

Задача 9.4. Компанія "XYZ"поставляє свою продукцію у 300 магазинів. Керівництво вирішило провести масштабну рекламну

кампанію, що коштувала 4 млн. грн. Щоб оцінити ефективність цієї рекламної кампанії було обстежено 30 магазинів ПВВбП, де компанія реалізує свою продукцію, до початку проведення рекламної кампанії та після. Сумарний виторг у обстежених магазинах до проведення рекламної кампанії (z_k) складав 1477.984 тис. грн, а після (y_k) – 1812.538 тис. грн.

1. Ефективність кампанії можна вимірювати різницею ($D = T_y - T_z$) сумарних виторгів *після* та *до* проведення рекламної кампанії. Оцінити параметр D та його дисперсію. Побудувати довірчий інтервал для D.

2. Також ефективність кампанії можна також вимірювати у процентному відношенні, тобто відношенням ($R = \frac{T_y}{T_z}$) сумарних виторгів *після* та *до* проведення рекламної кампанії. Оцінити параметр R та його дисперсію. Побудувати довірчий інтервал для R.

Що ви можете сказати про ефективність цієї рекламної кампанії?

Додаткова інформація: вибіркова дисперсія виторгів по магазинам до проведення кампанії $\widehat{S}_z^2 = 67.935$, після $\widehat{S}_y^2 = 109.696$, вибіркова коваріація між виторгами до і після $\widehat{S}_{yz} = 68.427$.

Задача 9.5. Вибірка 100 студентів була обрана ПВВбП із 1000. Нас цікавить чи здав студент екзамен чи ні, тобто є лише 2 результати: здав ("Успіх") та не здав ("Неуспіх"). Результати спостережень подані у таблиці

	Студенти (ч)	Студентки (ж)	Всього
Успіх	$n_{11} = 35$	$n_{12} = 25$	$n_{1.} = 60$
Неуспіх	$n_{21} = 20$	$n_{22} = 20$	$n_{2.} = 40$
Всього	$n_{.1} = 55$	$n_{.2} = 45$	$n = 100$

1. Оцініть рівень успішності (%) для студентів (R_1) та студенток (R_2).

2. Підрахуйте приблизне зміщення оцінки успішності для кожної із категорій (ч/ж).

3. Оцініть дисперсію оцінок рівня успішності та побудуйте 95% довірчі інтервали для кожного рівня успішності (ч/ж).

4. Проаналізуйте отримані довірчі інтервали та дайте відповідь на таке питання: "Чи можете ви з впевненістю сказати, грунтуючись на отриманій інформації, що успішність *студентів* вища за успішність *студенток*? Запропонуйте, що можна змінити, щоб мати змогу робити такі висновки?

Задача 9.6. Влада міста К вирішила фінансово допомагати сім'ям, що мають дітей віком до 1 року. Для визначення розміру доплат та загальних витрат на цю допомогу потрібно оцінити доходити сімей із малолітніми дітьми в розрахунку на одну особу. Відомо, що в цьому місті 10000 домогосподарств. А сімей, у яких є двоє дітей віком до 1 року немає. У місті проходило вибіркове обстеження, з якого стало відомо, що серед 30 обстежених ПВВбП домогосподарств у вибірку потрапили лише 3 домогосподарства із малолітніми дітьми у яких щомісячні доходи на 1 особу складають: 2, 5, та 3 тис. грн.

Середній дохід на одну особу для домогосподарств із малолітніми дітьми у цьому місті можна оцінити двома способами:

1. використати оцінку Горвіца-Томпсона $\widehat{\overline{y}}_d^{(HT)} = \frac{1}{n}\sum_{k\in s} y_k^*$, де

$$y_k^* = \begin{cases} y_k, & \text{k-те дг має малолітню динину;} \\ 0, & \text{k-те дг не має малолітніх дітей.} \end{cases}$$

2. використати альтернативну оцінку $\widehat{\overline{y}}_d^{(A)} = \frac{1}{n^*}\sum_{k\in s} y_k^*$, де n^* – це кількість дг, що мають малолітніх дітей у вибірці.

Як ви думаєте, яка із цих оцінок краща і в якому розумінні? Побудувати 95% довірчі інтервали для отриманих оцінок. Проаналізувати отримані результати.

Розділ 10

Використання допоміжної інформації

10.1. Основний теоретичний матеріал

Оцінювання за різницею.

Допоміжна змінна – це змінна, для якої наявна повна інформація до початку проведення обстеження.

Розглянемо J допоміжних змінних $x_1, \ldots, x_j, \ldots, x_J$. До початку вибіркового обстеження значення досліджуваної характеристики $y_1, \ldots, y_N$ залишаються невідомими, тоді як інформація про вектори $\mathbf{x}_1, \ldots, \mathbf{x}_N$ є у повному розпорядженні дослідника.

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 сумарне значення 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 – зазвичай у вигляді лінійної комбінації відомих значень допоміжних змінних $x_{1k}, \ldots, x_{Jk}$: $y_k^0 = \sum_{j=1}^J A_j x_{jk} = \mathbf{A}' x_k$, де $\mathbf{A} = (A_1, \ldots, A_J)'$ – вектор постійних коефіцієнтів.

Використовуючи наближені значення, невідоме сумарне значення можна записати у такому вигляді:

$$T_y = \sum_{k \in U} y_k = \sum_{k \in U} y_k^0 + \sum_{k \in U} (y_k - y_k^0) = \sum_{k \in U} y_k^0 + \sum_{k \in U} D_k. \qquad (10.1)$$

Наближене сумарне значення $\sum\limits_{k \in U} y_k^0$ у виразі (10.1) є відомою величиною, але сума різниць $\sum\limits_{k \in U} D_k$ є невідомою, тому природно замінити у формулі (10.1) суму різниць відповідною сумою

π-оцінок різниць:

$$\widehat{t}_{yD} = \sum_{k\in U} y_k^0 + \sum_{k\in s} \frac{D_k}{\pi_k}. \tag{10.2}$$

Такий метод називається *оцінювання за різницею* (англ. `difference estimator`). Оцінка $\widehat{t}_{yD}$ є незміщеною оцінкою сумарного значення $T_y = \sum\limits_{k\in U} y_k$. Її дисперсія обчислюється за формулою

$$\mathcal{D}(\widehat{t}_{yD}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{D_k}{\pi_k}\frac{D_l}{\pi_l}\,. \tag{10.3}$$

Незміщена оцінка дисперсії (10.3) обчислюється за формулою

$$\widehat{\mathcal{D}}(\widehat{t}_{yD}) = \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{D_k}{\pi_k}\frac{D_l}{\pi_l}\,, \tag{10.4}$$

де $\Delta_{kl} = (\pi_{kl} - \pi_k\pi_l)$, $k, l \in U : k \neq l$, та $\Delta_{kk} = \pi_k(1-\pi_k)$.

Якщо $y_k^0 = x_k$, то $D_k = y_k - x_k$. Тоді у випадку ПВВбП з розміром вибірки $n = fN$ оцінка за різницею набуде вигляду

$$\widehat{t}_{yD} = \sum_{k\in U} x_k + \sum_{k\in s} \frac{y_k - x_k}{n/N} = T_x + (\widehat{t}_{y\pi} - \widehat{t}_{x\pi})$$

з дисперсією

$$\mathcal{D}(\widehat{t}_{yD}) = N^2 \frac{1-f}{n}\left(S_y^2 + S_x^2 - 2S_{xy}\right), \tag{10.5}$$

яку можна оцінити як

$$\widehat{\mathcal{D}}(\widehat{t}_{yD}) = N^2 \frac{1-f}{n}\left(\widehat{S}_y^2 + \widehat{S}_x^2 - 2\widehat{S}_{xy}\right).$$

Узагальнена регресійна оцінка (GREG) базується на підході, подібному до оцінки за різницею

$$\widehat{t}_{yGREG} = \sum_{k\in U} y_k^0 + \sum_{k\in s} \frac{D_k}{\pi_k} = \widehat{t}_{y\pi} + \sum_{j=1}^{J} B_j(T_{x_j} - \widehat{t}_{x_j\pi}), \tag{10.6}$$

але значення коефіцієнтів в ній $B_1, \ldots, B_J$ є невідомими. Замість невідомих коефіцієнтів підставляють їх оцінки $\widehat{B}_1, \ldots, \widehat{B}_J$ для знаходження яких використовують моделі регресійного аналізу.

Оцінювання за відношенням. Нехай у нас є інформація про одну допоміжну змінну x, яка набуває тільки додатних значень.

Модель регресії M_1, яка базується на припущенні, що відношення y_k/x_k є приблизно константою, називається *моделлю відношення*. У цій моделі стверджується, що

$$\begin{cases} \mathrm{E}_{M_1}(y_k) = \beta x_k, \\ \mathcal{D}_{M_1}(y_k) = \sigma^2 x_k, \end{cases} \tag{10.7}$$

де значення параметрів β та σ^2 невідомі.

Регресійна оцінка, у якій наближені значення знаходяться виходячи із моделі (10.7), називається *оцінкою за відношенням* (англ. *ratio estimator*).

Оцінка за відношенням для T_y має такий вигляд:

$$\widehat{t}_{yR} = \left(\sum_{k \in U} x_k \right) \frac{\sum\limits_{k \in s} y_k/\pi_k}{\sum\limits_{k \in s} x_k/\pi_k} = T_x \, \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}} = T_x \, \widehat{B}. \tag{10.8}$$

Ця оцінка є приблизно незміщеною оцінкою параметра T_y, а наближене значення дисперсії оцінки $\widehat{t}_{yR}$ знаходиться за допомогою методу лінеаризації Тейлора за формулою

$$\widetilde{\mathcal{D}}(\widehat{t}_{yR}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{(y_k - B x_k)(y_l - B x_l)}{\pi_k \pi_l}, \tag{10.9}$$

де

$$B = \frac{\sum\limits_{k \in U} y_k}{\sum\limits_{k \in U} x_k} = \frac{T_y}{T_x}.$$

Оцінка цієї дисперсії

$$\widehat{\mathcal{D}}(\widehat{t}_{yR}) = \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \left(\frac{g_{ks}(y_k - \widehat{B} x_k)}{\pi_k} \right) \left(\frac{g_{ls}(y_l - \widehat{B} x_l)}{\pi_l} \right) = \tag{10.10}$$

де для всіх $k \in s$

$$g_{ks} = \frac{\sum\limits_{k \in U} x_k}{\sum\limits_{k \in s} x_k/\pi_k} = \frac{T_x}{\widehat{t}_{x\pi}}. \tag{10.11}$$

Оцінювання за лінійною регресією.

Модель регресії M_2, яка базується на припущенні, що залежність між y_k та однією допоміжною змінною x_k є лінійною виду $y_k \approx \beta_1 + \beta_2 x_k$, називається *лінійною регресійною моделлю*. Тобто, в даному випадку регресійна пряма *не* обов'язково проходить через початок координат. Модель лінійної регресії має вигляд:

$$\begin{cases} \mathrm{E}_{M_2}(y_k) = \beta_1 + \beta_2 x_k, \\ \mathcal{D}_{M_2}(y_k) = \sigma^2, \quad \forall k, \end{cases} \tag{10.12}$$

де значення параметрів β_1, β_2 та σ^2 невідомі.

Узагальнена регресійна оцінка (GREG), яка ґрунтується на лінійній моделі (10.12), називається *оцінкою за лінійною регресією* (англ. `linear regression estimator`).

Коефіцієнти цієї моделі можна оцінити наступним чином:

$$\widehat{B}_1 = \frac{\sum_{k \in s} \frac{1}{\pi_k}(x_k - \frac{\widehat{t}_{x\pi}}{\widehat{N}})(y_k - \frac{\widehat{t}_{y\pi}}{\widehat{N}})}{\sum_{k \in s} \frac{1}{\pi_k}(x_k - \frac{\widehat{t}_{x\pi}}{\widehat{N}})^2}; \tag{10.13}$$

$$\widehat{B}_0 = \widehat{\overline{y}}_\pi - \widehat{B}_1 \widehat{\overline{x}}_\pi, \quad \text{де} \quad \widehat{N} = \sum_{k \in s} \frac{1}{\pi_k}. \tag{10.14}$$

Оцінка за лінійною регресією для сумарного T_y має вигляд:

$$\begin{aligned} \widehat{t}_{yLR} &= \sum_{k \in U} \widehat{y}_k + \sum_{k \in s} \frac{y_k - \widehat{y}_k}{\pi_k} = \\ &= \sum_{k \in U} (\widehat{B}_0 + \widehat{B}_1 x_k) + \sum_{k \in s} \frac{y_k - \widehat{B}_0 - \widehat{B}_1 x_k}{\pi_k} = \\ &= \widehat{t}_{y\pi} + \widehat{B}_0(N - \widehat{N}) + \widehat{B}_1(T_x - \widehat{t}_{x\pi}). \end{aligned}$$

Якщо для вибіркового дизайну $p(s)$: $\widehat{N} = N$ (як наприклад, для ПВВбП), тоді

$$\widehat{B}_1 = \frac{\sum_{k\in s}\frac{1}{\pi_k}(x_k - \widehat{\overline{x}}_\pi)(y_k - \widehat{\overline{y}}_\pi)}{\sum_{k\in s}\frac{1}{\pi_k}(x_k - \widehat{\overline{x}}_\pi)^2} = \frac{\widehat{S}_{xy}}{\widehat{S}_x^2}.$$

Ця оцінка також є приблизно незміщеною оцінкою параметра T_y при великих N та n.

Наближене значення дисперсії оцінки $\widehat{t}_{yLR}$ можна отримати методом ліанеризації Тейлора. Формула має вигляд:

$$\widetilde{\mathcal{D}}(\widehat{t}_{yLR}) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{(y_k - B_0 - B_1x_k)(y_l - B_0 - B_1x_l)}{\pi_k\pi_l}, \tag{10.15}$$

де

$$B_1 = \frac{\sum_{k\in U}(x_k - \overline{X})(y_k - \overline{Y})}{\sum_{k\in U}(x_k - \overline{X})^2} = \frac{S_{xy}}{S_x^2},$$

$$B_0 = \overline{Y} - B_1\overline{X}.$$

Оцінка дисперсії оцінки $\widehat{t}_{yLR}$

$$\widehat{\mathcal{D}}(\widehat{t}_{yLR}) = \sum_{k\in s}\sum_{l\in s}\frac{(\pi_{kl} - \pi_k\pi_l)}{\pi_{kl}}\cdot\frac{(y_k - \widehat{B}_0 - \widehat{B}_1x_k)}{\pi_k}\cdot\frac{(y_l - \widehat{B}_0 - \widehat{B}_1x_l)}{\pi_l}. \tag{10.16}$$

Зокрема, при ПВВбП оцінка за лінійною регресією буде мати вигляд $\widehat{t}_{yLR} = N\overline{y} + \frac{\widehat{S}_{xy}}{\widehat{S}_x^2}(\overline{X} - \overline{x})$, а її дисперсія буде оцінюватися за формулою:

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

де $\widehat{S}_e^2$ – це вибіркова дисперсія залишків лінійної регресії $e_k = y_k - \widehat{B}_0 - \widehat{B}_1x_k$.

10.2. Задачі

Задача 10.1. 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.

Задача 10.2. Із генеральної сукупності розміру $N = 200$ вибіркових одиниць за допомогою ПВВбП 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 дисперсію цієї оцінки.

Задача 10.3. Припустимо, що в місті Статислав потрібно оцінити сумарну кількість осіб віком 60+. Для цього серед 1024 домогосподарств (дг) було вибрано ПВВбП та обстежено 20 дг. Результати обстеження наступні: лише в 3 дг проживали особи віком 60+, при чому у 2-х дг проживало по двоє таких осіб, а у одному дг проживала одна особа віком 60+. Всі інші дг не мали в своєму складі осіб віком 60+.

Оцінити шукану кількість та дисперсію цієї оцінки:

1) використовуючи оцінку Горвіца-Томпсона;

2) використовуючи оцінку за різницею;

3) використовуючи оцінку за відношенням;

4)використовуючи оцінку за лінійною регресією;

якщо відомо із зовнішніх джерел, що всього в Статиславі домогосподарств, що мають осіб віком 60+ є 300. Порівняти отримані результати. Який метод оцінювання виявився найбільш точним?

Задача 10.4. Влада міста Т вирішила фінансово допомагати сім'ям, що мають дітей віком до 1 року. Для визначення сумарного розміру витрат на цю допомогу провели вибіркове обстеження в якому дослідили доходити сімей із малолітніми дітьми в розрахунку на одну особу. З нього стало відомо, що серед 100 обстежених ПВВбП домогосподарств у вибірку потрапили лише 3 домогосподарства із малолітніми дітьми, що мають такі щомісячні доходи на особу: 2, 5, та 3 тис. грн.

Відомо, із адміністративних джерел, що в цьому місті є 10000 домогосподарств, 108 із яких мають малолітніх дітей. При цьому сімей, у яких є двоє дітей віком до 1 року немає.

Доплати для таких сімей будуть розраховуватись наступним чином: якщо дохід на одну особу не перевищує 2500 грн, тоді доплата буде становити різницю між цією сумою та доходом на одну особу для цього домогосподарства. Якщо дохід є вищим за цю суму, тоді доплату робити не будуть.

Оцінити загальну суму доплат для сімей із дітьми в цьому місті та оцінити дисперсію для цієї величини використовуючи:

1. оцінку *Горвіца-Томпсона*;

2. оцінку *за різницею*. В якості допоміжної змінної можна взяти змінну $x_k = 200$ грн, якщо дг має малолітніх дітей, та $x_k = 0$ – якщо немає;

3. оцінку *за відношенням*. В якості допоміжної змінної можна взяти змінну $x_k = 1$, якщо дг має малолітніх дітей, та $x_k = 0$ – якщо немає. Якою буде в цьому випадку оцінка коефіцієнта відношення $\widehat{B}$?

4. оцінку *за лінійною регресією*. В якості допоміжної змінної використайте змінну із попереднього пункту.

Яка із цих оцінок краща і в якому розумінні? Побудувати 95% довірчі інтервали для отриманих оцінок.

Задача 10.5. Два лікаря-дантиста проводять обстеження стану зубів у 200 дітей в одному населеному пункті. Перший дантист вибрав простим випадковим відбором без повернення 20 дітей із 200 та зібрав дані про кількість зубів із карієсом для кожної дитини із вибірки. Отримані дані наведені в таблиці.

Кількість зубів із карієсом	0	1	2	3	4	5	6	7	8
Кількість дітей	8	4	2	2	1	2	0	0	1

Інший дантист обстежив всіх 200 дітей, але його цікавила лише кількість дітей, що взагалі не мають карієса. Таких виявилось 50.

1. Оцінити середню кількість зубів уражених карієсом на одну дитину використовуючи лише інформацію, що зібрав перший дантист. Оцініть дисперсію цієї оцінки та побудуйте 95% довірчий інтервал.

2. Оцінити середню кількість зубів уражених карієсом на одну дитину використавши також інформацію, що зібрав другий дантист. Оцініть дисперсію цієї оцінки. Яким є ефект від використання додаткової інформації?

3. Поміркуйте, чи отриманий ефект від використання додаткової інформації вартий залучення ще одного дантиста до збору інформації?

Розділ 11

Лабораторні роботи

11.1. Лабораторні роботи на основі гіпотетичної популяції людей, що проживають в області Стефенс

Опис

Гіпотетична область Стефенс була розроблена Т. Чангом та Ш.Лор для навчальних цілей. Припускається, що ця гіпотетична область Стефенс, що розташована в США. Дані для обстежень отримуються за допомогою програми `SURVEY`. Матеріали та зображення було взято із книги Ш.Лор [18].

Область Стефенс поділена на 75 районів (districts), нумерація будинків в кожному з районів почитається з 1. Мапа районів області Стефенс зображена на Рис. 11.1..

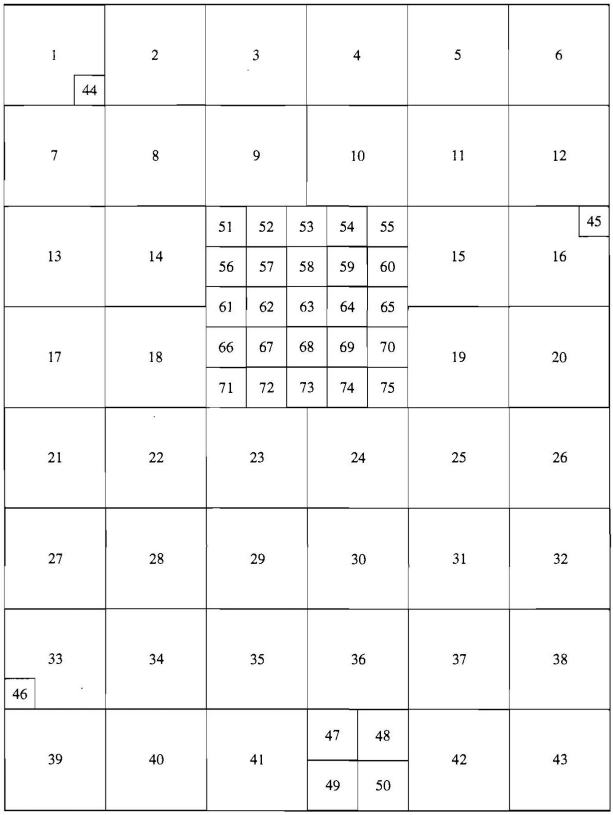
Область Стефенс має населення близько 103 тис. чоловік. Вона має 2 основних великих міста: Локхард (райони 51-75) з населенням 57 500 чол. та Івествіль (райони 47-50) з населенням 11 700 чол. Обида міста є транспортними та комерційними центрами, мають виробництво. Також область містить 3 менші населені пункти: містечка Вілегас (район 44), Велдон (район 45) та Роутледж (район 46), з населенням від 1 000 до 2 000 чол. Населення цих міст та містечок вважається міським населенням. Населення, що проживає поза цими містами (райони 1-43) — сільським населенням. Більш докладна інформація про кількісний склад районів — на Рис. 11.1..

Кабельна компанія SCCC планує виходити на ринок області Стефенс і перед початком кампанії проводить дослідження, що направлене на визначення цінової політики та програмного наповнення своїх послуг. Для цього вона проводить вибіркове обстеження домогосподарств.

В обстеженні використовується Анкета з наступними вмістом:

Вітаємо Вас! Ми проводимо обстеження для кабельної компанії SCCC. Відповівши на наші питання Ви допоможете нам зробити наш сервіс якомога корисним для вас. Дякуємо за спів-

Рис. 11.1: Мапа районів області Стефенс



працю!

1. *Скільки осіб віком 12 років і більше проживають за вашою адресою? (Будь-ласка, включайте лише осіб, які ви можете вважати членами своєї родини. Не включайте осіб, що винаймають у вас кімнату або є гостями.)*

2. *Скільки осіб віком 11 років і менше проживають за вашою адресою?*

3. *Скільки телевізорів є в вашому помешканні?*

Рис. 11.2: Інформація про кількісний склад районів: (1) номер району; (2) кількість будинків в районі; (3) акумульовані кількості будинків; (4) кількість людей, що проживають в районі; (5) середня вартість будинку в районі.

(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
1	142	142	526	65248.	39	95	7390	312	57174.
2	153	295	624	58759.	40	130	7520	446	55702.
3	135	430	508	62319.	41	152	7672	533	53285.
4	128	558	560	59416.	42	169	7841	672	56866.
5	110	668	455	57202.	43	91	7932	371	50710.
6	103	771	404	59290.	44	283	8215	1029	60057.
7	105	876	421	71122.	45	562	8777	2079	57233.
8	385	1261	1488	79265.	46	312	9089	1149	52719.
9	296	1557	1112	75921.	47	897	9986	3263	62034.
10	287	1844	994	68254.	48	734	10720	2623	60764.
11	253	2097	929	60660.	49	963	11683	3490	60010.
12	172	2269	628	53569.	50	642	12325	2318	54498.
13	198	2467	768	65182.	51	525	12850	1825	95123.
14	432	2899	1595	77907.	52	726	13576	2497	68406.
15	248	3147	864	65739.	53	674	14250	1948	53634.
16	251	3398	915	53771.	54	585	14835	1219	48643.
17	221	3619	864	68257.	55	553	15388	1090	43493.
18	297	3916	1099	78449.	56	583	15971	1977	95110.
19	235	4151	812	70772.	57	911	16882	2691	84394.
20	171	4322	687	52711.	58	1051	17933	2663	57657.
21	135	4457	525	66739.	59	918	18851	1824	36706.
22	254	4711	923	66249.	60	799	19650	1636	44308.
23	203	4914	708	74757.	61	545	20195	1853	101906.
24	244	5158	825	75766.	62	895	21090	2588	74815.
25	202	5360	799	68989.	63	1313	22403	2642	55560.
26	103	5463	388	56994.	64	968	23371	2457	62813.
27	102	5565	398	58940.	65	717	24088	2203	69846.
28	115	5680	448	60448.	66	651	24739	2197	93771.
29	180	5860	693	69111.	67	886	25625	2711	82902.
30	190	6050	766	69685.	68	912	26537	2750	76832.
31	152	6202	633	70276.	69	898	27435	2671	72062.
32	141	6343	572	63819.	70	759	28194	2650	79887.
33	143	6486	610	58636.	71	722	28916	2568	87383.
34	135	6621	491	55554.	72	753	29669	2652	80341.
35	178	6799	699	62361.	73	793	30462	2763	79833.
36	221	7020	811	60052.	74	725	31187	2560	83354.
37	174	7194	719	59699.	75	802	31989	2870	80522.
38	101	7295	390	53322.					

4. *Якщо послуги кабельного телебачення будуть коштувати 5$ в місяць, чи ви б скористались його послугами? А якщо 10$, 15$, 20$...? (Інтерв'юер фіксує найвищу ціну, що погоджуються заплатити на послуги КТБ).*

5. *Скільки годин Ви особисто провели біля телевізора впродовж останнього тижня? Ваш чоловік(дружина)? Діти? Інші особи, що проживають з Вами? (Інтерв'юер записує загальну суму всіх годин. Якщо є присутні вдома інші члени родини, питання їм ставиться окремо).*

6. *Скільки годин ви дивились новини та(чи) інші публіцистичні програми?*

7. *Скільки годин ви дивились спортивні програми?*

8. *Скільки годин ви дивились дитячі програми?*

9. *Скільки годин ви дивились художні фільми?*

Додатково, кабельна компанія отримала інформацію із реєстру будинків про оціночну вартість всіх житлових будинків в області Стефенс. Ця інформація відома про всі будинки і не запитується під час інтерв'ю.

Витрати на проведення обстеження:

- 60$ витрачаються на транспортні витрати для поїздок в райони сільської місцевості (1-46)
- 20$ витрачаються на транспортні витрати для поїздок в райони міської місцевості (47-50)
- 6$ витрачаються на відвідування та опитування домогосподарств в районах сільської місцевості (1-46, в незалежності від того, був хтось опитаний чи ні)
- 3$ витрачаються на відвідування та опитування домогосподарств в міських районах місцевості (включаючи м. Локхард, в незалежності від того, був хтось опитаний чи ні)
- 10$ витрачаються на опрацювання кожного отриманого інтерв'ю.

Наприклад, якщо в вибірку потрапили адреси: 3–47 (3 район, 47), 3–25, 5–16, 51–25, 51–36, то вартість такого обстеження буде обраховуватись так: 2*60+1*20+3*6+2*3+5*10=214 $.

Лабораторна робота № 1. ПВВбП

1. Провести пробне обстеження 20 домогосподарств в області Стефенс за допомогою простого випадкового відбору без повернення. Оцінити:

   (a) загальну кількість телевізорів в області Стефенс;

(б) середню ціну, яку погоджуються платити за послуги кабельного телебачення;

(в) дисперсії генеральної сукупності (S^2) по характеристиці y – «кількість телевізорів» та по характеристиці z – «ціна за послуги кабельного ТБ».

2. Підрахувати необхідний розмір вибірки, використовуючи інформацію з пробного обстеження, для проведення повноцінного обстеження, якщо потрібно оцінити одночасно та з 95% вірогідністю: а) загальну кількість телевізорів в області Стефенс з *абсолютною* точністю $= 2000$; б) середню ціну, яку погоджуються платити за послуги кабельного телебачення з *відносною* точністю $\widetilde{e} = 2\%$. Чи достатньо буде для отримання бажаних точностей вибірки розміру 200?

3. Провести обстеження 200 домогосподарств та побудувати 95% довірчі інтервали для: а) загальної кількості телевізорів в області Стефенс; б) середньої ціни, яку погоджуються платити за послуги кабельного телебачення; в) частки домогосподарств, які погоджуються платити за кабельне ТБ хоча б 10 \$. Якого типу передачі найбільш популярні в області Стефенс?

4. По даним з вибірки побудувати 95% довірчий інтервал для середньої вартості будинків в області Стефенс. Чи містить отриманий інтервал реальне значення 68 045 \$? На вашу думку, чи отримана вибірка є репрезентативною? Чому?

Лабораторна робота № 2 Систематичний відбір

1. Для обстеження жителів області Стефенс застосувати систематичний відбір з розміром вибірки $n = 200$. Яким буде вибірковий інтервал a в цьому випадку? Оцінити:

   (а) загальну кількість телевізорів в області Стефенс;

   (б) середню ціну, яку погоджуються платити за послуги кабельного телебачення;

(в) частку домогосподарств, які погоджуються платити за кабельне ТБ хоча б 10 $.

Оцінити відповідні дисперсії за допомогою зміщених оцінок $\hat{D}$. Обгрунтувати доцільність та ризики їх використання.

2. Застосувати для обстеження систематичний відбір з 2-ма випадковими стартами ($m = 2$) та вибірковим інтервалом ma, так щоб розмір загальної вибірки співпадав з розміром вибірки з п.1. Оцінити:

   (а) загальну кількість телевізорів в області Стефенс;

   (б) середню ціну, яку погоджуються платити за послуги кабельного телебачення;

   (в) частку домогосподарств, які погоджуються платити за кабельне ТБ хоча б 10 $

   та відповідні дисперсії. Порівняти з результатами з п.1. та лабораторної роботи 1 (п.3). Проаналізувати.

Лабораторна робота №3. Відбір із поверненням

1. Провести обстеження домогосподарств в області Стефенс за допомогою простого випадкового відбору з поверненням з розміром вибірки 200. Оцінити:

   (а) загальну кількість телевізорів в області Стефенс;

   (б) середню ціну, яку погоджуються платити за послуги кабельного телебачення;

   (в) відсоток тих домогосподарств, що мають дітей.

   Вказати для отриманих оцінок довірчі інтервали.

2. Провести обстеження домогосподарств в області Стефенс за допомогою відбору з поверненням з ймовірностями пропорційними до вартості будинку домогосподарства з розміром вибірки 200. Оцінити в цьому випадку:

(а) загальну кількість телевізорів в області Стефенс за допомогою оцінки Хансена-Гурвіца;

(б) загальну кількість телевізорів в області Стефенс за допомогою оцінки Горвіца-Томпсона.

Яка з цих оцінок краща? Аргументуйте відповідь.

Лабораторна робота №4. Стратифікований відбір

1. Провести обстеження ПВВбП з метою визначення середньої ціни, яку погоджуються платити за послуги кабельного телебачення в області Стефенс з розміром вибірки n=200. Оцінити дисперсію цієї оцінки. Яка вартість проведення такого обстеження? (Тут можна просто використати результати, отримані в Лабораторній роботі №1).

2. Населення області Стефенс можна поділити на страти двома способами:

   1) Сільське населення (райони 1—46) та міське населення (райони 47—75);
   2) Сільське населення (райони 1—43), міське населення крім міста Локхард (райони 44—50) та місто Локхард (райони 51—75).

3. Яким буде пропорційне розміщення при простому стратифікованому відборі для вибірки розміру 200 для цих двох випадків? Оцінити середню ціну, яку погоджуються платити за послуги кабельного телебачення в області. Оцінити дисперсію цієї оцінки. Порівняти з п.1). Яка вартість проведення таких обстежень?

4. Використовуючи значення вибіркової дисперсії з п.2) та, припускаючи, що вартість обстеження в усіх стратах однакова, знайти розміщення Неймана для стратифікованої вибірки розміру 200 для цих двох способів стратифікації. Оцінити середню ціну, яку погоджуються платити за послуги кабельного телебачення в області Стефенс. Оцінити дисперсію цієї оцінки для обох видів стратифікації. Порівняти з п.1) та 2).

5. Вважаючи, що для районів 1—46 вартість обстеження 1 домогосподарства складає 76$, а для районів 47—75 вартість обстеження 1 домогосподарства складає 33$ підрахувати оптимальне розміщення при фіксованих витратах C=10 000$ для першого способу стратифікації використовуючи значення вибіркової дисперсії з п.2). Оцінити середню ціну, яку погоджуються платити за послуги кабельного телебачення в області. Якою буде мінімальна дисперсія цієї оцінки? Порівняти з п.1) 2) та 3).

6. Проаналізувати отримані результати. Можливо ви можете запропонувати якусь альтернативний спосіб стратифікацію? Напишіть свої міркування.

Лабораторна робота № 5. Кластерний відбір

Потрібно перевірити чи кластерна вибірка області Стефенс може бути кращою (точнішою) за просту випадкову вибірку без повернення розміру 200 при однаковій вартості. Припускається, що основним параметром, що цікавить кабельну компанію, є загальна кількість телевізорів в області Стефенс. Для цього:

1. потрібно оцінити вартість обстеження 200 домогосподарств ПВВбП використавши програму `ADDGEN` для отримання 10 різних вибірок ПВВбП. Усереднивши вартість отримання 10 ПВВбП вибірок ми отримаємо оцінку вартості проведення такого обстеження. Якою буде ця середня вартість?

2. спланувати двостадійну кластерну вибірку таким чином, щоб отримати найменшу дисперсію для оцінки загальної кількості телевізорів в області Стефенс при тій же середній вартості обстеження як і вартість ПВВбП вибірки з п. 1 (для визначення необхідних параметрів можна скористатись інформацією щодо загальної кількості телевізорів в області Стефенс взятих із попередніх обстежень (лабораторних робіт)). Скільки потрібно буде обстежити домогосподарств в цьому випадку?

3. Провести обстеження домогосподарств використовуючи двостадійний кластерний відбір із параметрами, визначеними в п. 2. Порахувати:

   (а) оцінку загальної кількості телевізорів в області Стефенс;

   (б) оцінку середньої ціни, яку погоджуються платити за послуги кабельного телебачення;

   (в) вказати для отриманих оцінок оцінки дисперсії та довірчі інтервали. Порівняти з дисперсією відповідних оцінок при ПВВбП розміру 200.

11.2. Лабораторні роботи на основі гіпотетичної популяції людей, що проживають на Островах

Опис

Цю штучну віртуальну популяцію людей, що проживають островах, було розроблено професорами Майклом Балмером та Кімберлі Халадін з Університету Квінсленда (Австралія) для навчання та викладання експериментального дизайну, епідеміології та інших статистичних досліджень. Більш докладно про цей ресурс написано в [12].

Острів (The Island) можна знайти за адресою http://island.maths.uq.edu.au. Цей ресурс безкоштовний, але для можливості ним користуватись треба отримати логін та пароль. Для цього викладач може звернутись за адресою island@maths.uq.edu.au. Студентів залучає до цього ресурсу вже безпосередньо викладач.

Жителі острова живуть у двадцяти семи поселеннях на трьох островах. На головній карті ви можете натиснути на поселення, щоб відвідати його. Кожне поселення складається з багатьох будинків, а також кількох інших будівель (наприклад, школи чи адміністрації). При натисканні на них, ви побачите сім'ю, яка живе в ньому. Натисніть на одне з імен у родині, щоб детально ознайомитися з жителями острова. Кожен житель острова розповідає вам свою історію, якої часто достатньо для обстеження.

Особливістю цього віртуального ресурсу є те, що особи в ньому мають дуже наближену поведінку до реальної. Вони не відповідають на запитання вночі, коли сплять. Або можуть взагалі не захотіти відповідати - таким чином ми можемо отримувати невідповіді. Населення остовів досить мінливе - вони переселяються в інші поселення, помирають, народжуються. Це ускладнює роботу із списком генеральної сукупності, що потрібен при проведенні вибіркового обстеження. Але це також показує важливість роботи із побудовою і підтримкою в актальному стані такого списку.

Лабораторна робота № 1. ПВВбП

1. Провести пробне обстеження Островів 10 осіб працездатного віку за допомогою простого випадкового відбору без повернення. Оцінити :

   - сумарний дохід (T) працездатного населення Островів;
   - середній рівень IQ ($\overline{Y}$) для працездатного населення;
   - пропорцію тих, хто є безробітним (P, у відсотках);
   - дисперсії генеральної сукупності (S^2) по всім трьом характеристикам;
   - дисперсії оцінок всіх трьох параметрів $\widehat{\mathcal{D}}(\widehat{\theta})$.

2. Підрахувати необхідний розмір вибірки для проведення обстеження населення островів працездатного віку (можна використовувати інформацію з пробного обстеження), для проведення повноцінного обстеження, якщо потрібно оцінити всі параметри ОДНОЧАСНО з 95% вірогідністю та точністю, що задана окремо для кожного параметра:

   - сумарний дохід працездатного населення із відносною точністю 20%;
   - середній рівень IQ для працездатного населення із абсолютною точністю 5 одиниць;
   - пропорцію тих, хто є безробітним (у відсотках) із абсолютною точністю 1,5%;

3. Провести вибіркове обстеження всього населення островів працездатного віку за допомогою вибірки на основі ПВВбП для 30 осіб, оцінити три параметри та їх дисперсії, побудувати 95% довірчі інтервали для всіх трьох параметрів;

4. Порахувати по отриманій вибірці із п. 2) середній вік осіб та побудувати для цього параметру 95% довірчий інтервал. Чи входить у отриманий довірчий інтервал реальний середній вік 39.27? Який висновок ви можете із цього зробити?

Лабораторна робота № 2. Стратифікований відбір

Працездатне населення Островів поділити на страти хоча б двома способами. Змінні, які можна використати для стратифікації: стать, вік та місце проживання.

1. Яким буде пропорційне розміщення при простому стратифікованому відборі для вибірки розміру 30 для цих двох випадків? Оцінити сумарний дохід працездатного населення Островів, середнє IQ та пропорцію безробітних. Оцінити дисперсію цієї оцінки.

2. Використовуючи значення вибіркової дисперсії з п.1) та припускаючи, що вартість обстеження у всіх стратах однакова, знайти розміщення Неймана для стратифікованої вибірки розміру 30 для обох видів стратифікації. Оцінити сумарний дохід працездатного населення Островів, середнє IQ та пропорцію безробітних для розміщення Неймана. Оцінити дисперсію цієї оцінки.

3. Проаналізувати отримані результати — можна порівняти отримані результати із тими, що були отримані при ПВВбП в лаб.роботі № 1 та в п.1) та п. 2). Який із методів оцінювання виявився найкращим? Чи можете ви сказати який із островів найбагатший, де населення має більший рівень IQ, де найбільший рівень безробіття? Напишіть свої міркування.

11.3. Лабораторні роботи на основі гіпотетичного селища Статвілідж

Опис

Статвілідж (StatVillage) — це гіпотетичне поселення в Канаді. Знайти його можна за адресою http://jse.amstat.org/v5n2/schwarz.supp/index.html і працювати можливо прямо в веб-браузері. Розробником цього ресурсу є професор Карл Дж. Шварц із Університету Саймона Фрезера (Канада) [21].

Будинки в Статвіліджі упорядковані системою блоків на прямокутній сітці з 8 будинками в одному блоці. Посередині кожної групи з 8 будинків є ігровий майданчик. Адреси будинків формуються за допомогою блочної системи. Послуги (наприклад, продовольчі магазини, магазини) розташовані на периферії поселення і не відображаються на карті. Домогосподарства можна вибрати для опитування за допомогою клікабельної карти.

Є три версії поселення: максимальне поселення - 128 кварталів, міні поселення - 60 кварталів та мікропоселення - 36 кварталів. Результати опитування повертаються у ваш веб-браузер, а потім можуть бути збережені у файл, наприклад, `.txt` для подальшої обробки.

Для кожного домогосподарства вимірюється багато різних змінних, пояснення про кожну міститься в кодовій книзі, що пояснює різні змінні та їх коди. Як і в більшості міст, деякі райони міста є більш заможними, ніж інші райони. У цьому поселенні люди з вищими доходами, в основному, зосереджені у верхній частині села, тоді як люди з низькими доходами, зосереджені у нижній частині села.

Дані, отримані для кожного домогосподарства, є реальними даними, взятими з файлів мікроданих загального користування з перепису населення Канади 1991 року.

При роботі із Статвіліджем студентам пропонується придумати для себе свою проблему, яка буде досліджуватись за допомогою вибіркового обстеження. Для цього їм пропонується вибрати із доступних змінних ті, які будуть аналізуватись різними пара-

метрами: сумарним, середнім, пропорцією. Студенти можуть працювати парами.

Приклад задачі для обстеження. Сім'я з двома дітьми планує з великого мегаполісу переїхати в маленьке затишне містечко. Їх вибір пав на Статвілідж. Але перед тим, як переїжджати, вони хочуть дізнатись певну інформацію про це місто. Які питання їх цікавлять?

1. Скільки дітей проживає в Статвіліджі?(сумарне)
2. Яка середня вартість житла в Статвіліджі? (середнє)
3. Яка частка домогосподарств здається в аренду? (пропорція)

Змінні, що будуть досліджуватись в цій задачі: для сумарного: HHPERA (діти до 5-ти років); для середнього: VALUEH; для пропорції: TENURH.

Важливим є здобуття студентами навичок не лише робити технічні підрахунки, а й аналізу отриманих результатів. Тому обов'язково в кінці кожної лабораторної роботи повинен бути висновок, що стосується поставленою студентами самим собі задачі. Так у наведеному прикладі, потрібно вирішити, чи буде сім'я переїзджати у Статвілідж чи ні.

Лабораторна робота № 1. ПВВбП

1. Провести пробне обстеження домогосподарств Статвіліджа (максимальне поселення - 128 blocks) із часткою відбору f=n/N=2% за допомогою простого випадкового відбору без повернення. Оцінити для своєї задачі:
   - сумарне T — по одній змінній $y_k^{(1)}$;
   - середнє $\overline{Y}$ – по іншій змінній $y_k^{(2)}$;
   - пропорцію P — по ще одній змінній $y_k^{(3)}$;
   - дисперсії генеральної сукупності ($S^2_{y^{(i)}}$) по всім трьом характеристикам $i = 1, 2, 3$, що будуть використовуватись в обстеженні.

- дисперсії оцінок всіх трьох параметрів ($\widehat{\mathcal{D}}(\widehat{\theta})$, $\widehat{\theta} \in \{\widehat{t}, \widehat{\overline{y}}, \widehat{p}\}$).

2. Підрахувати необхідний розмір вибірки, використовуючи інформацію з пробного обстеження, для проведення повноцінного обстеження, якщо потрібно оцінити ОДНОЧАСНО: параметри сумарне T та середнє $\overline{Y}$ з 95% вірогідністю та відносною похибкою 20%, а параметр пропорція P (вимірюється у %) – із абсолютною допустимою похибкою 10% та 95% вірогідністю.

3. Провести повноцінне обстеження тієї кількості домогосподарств, що були отримані в п. 2 та побудувати 95% довірчі інтервали для тих трьох параметрів, що цікавлять в обстеженні.

4. Порівняти отримані результати в п. 1 та п. 3: як вони змінились, на скільки сильно, чому це сталось?

5. Який висновок ви можете зробити стосовно проблеми, яку ви досліджували?

Лабораторна робота № 2. Оцінювання відношення

1. Виберіть для своєї задачі параметр, що можна представити як відношення R (наприклад, R це може бути дохід на одну особу, зокрема цей параметр отримуємо якщо поділити загальний дохід домогосподарства на кількість осіб у домогосподарстві). Провести обстеження 1/4 частини домогосподарств Статвіліджа (128 блоків) за допомогою простого випадкового відбору без повернення. Оцінити параметр R.

2. Оцінити дисперсію оцінки $\widehat{R}$ трьома методами: за допомогою лінеаризації Тейлора, Джек-найф та Бут-стреп на основі однієї і тієї ж вибірки.

3. Побудувати 95% довірчі інтервали для параметра R, використовуючи оцінки для дисперсії з п. 2 а також використовуючи вибірковий розподіл оцінки у випадку використання методу Бутстреп. Проаналізувати отримані результати.

Лабораторна робота № 3. Використання допоміжної інформації

В Статвіліджі потрібно оцінити середній дохід д/г, (змінна y – це `TOTINCH`, вважається відомою лише для елементів із вибірки) використовуючи як допоміжну, інформацію про:

а) кількість осіб, що проживають в д/г (змінна `HHSIZE`, вважається відомою для всіх дг);

б) вартість будинку (змінна `VALUEH`, вважається відомою для всіх дг).

1. Провести обстеження половини домогосподарств Статвіліджа (128 блоків) за допомогою простого випадкового відбору без повернення. Оцінити середнє змінної y використовуючи оцінку за відношенням та оцінку за лінійною регресією з використанням кожної із допоміжних змінних із п. а) та б).

2. Оцінити дисперсії цих оцінок на основі однієї і тієї ж вибірки. Проаналізувати отримані результати. На скільки точними є отримані результати порівняно одна із іншою та із оцінкою Горвіца-Томпсон?

3. Хоча б для одного випадку оцінити дисперсію одним із альтернативних методів — джек-найф чи бут-стреп. Порівняти із результатом, що був отриманий методом ліанеризації Тейлора (класичний).

4. Побудувати 95% довірчі інтервали для цих оцінок.

Фінальний проект

Провести одне обстеження тієї кількості домогосподарств Статвіліджа (максимальне поселення - 128 блоків), що ви отримали в п. 2 Лабораторної роботи № 1 за допомогою ПВВбП. (можна працювати із тією ж самою вибіркою, що була отримана в п. 3 Лабораторної роботи № 1). Для змінних та параметрів із лабораторної роботи № 1:

- сумарне — одна змінна $y_k^{(1)}$, вважається відомою лише для елементів із вибірки, та невідомою для всіх інших елементів генеральної сукупності;
- середнє – інша змінна $y_k^{(2)}$, вважається відомою лише для елементів із вибірки, та невідомою для всіх інших елементів генеральної сукупності;
- пропорцію (частка) — ще одна дихотомічна змінна $y_k^{(3)}$, вважається відомою лише для елементів із вибірки, та невідомою для всіх інших елементів генеральної сукупності

Виписати:

1. оцінки параметрів із п. 1-3 за допомогою оцінки Горвіца-Томспона разом із оцінками для їх дисперсій;
2. оцінки параметрів із п. 1-3 за допомогою оцінювання за відношенням вибравши для цього допоміжну змінну (чи декілька змінних) x_k, що буде вважатись відомою для всіх без винятку елементів генеральної сукупності (домогосподарств Статвіліджа). Для різних змінних можна вибирати різні допоміжні змінні (хоча і не обов'язково), враховуючи, що вони повинні бути пов'язані із змінною, що досліджується. Порахувати оцінки дисперсій для оцінок за відношенням для цих трьох параметрів;
3. оцінки параметрів із п. 1-3 за допомогою оцінювання за лінійною регресією взявши для цього ту ж саму допоміжну змінну(ні) x_k, що і в п. 2), що буде вважатись відомою для всіх без винятку елементів генеральної сукпності. Порахувати оцінки дисперсій для оцінок за лінійнійною регресією для цих трьох параметрів.
4. Проаналізувати отримані результати — коли та чому ви отримали найкращий результат? (Бажано зібрати всі результати в одній таблиці). Які будуть ваші рекомендації? Які висновки ви зробите стосовно тієї проблеми, що ви з самого початку збирались вирішувати?

Розділ 12

Відповіді та розв'язки

Відповіді та вказівки до розділу 1

1.1 *Вказівка.* Позначимо тут множину всіх осіб із яких робиться відбір як U_1 та вибірку осіб як s_1. Тоді ймовірність включення особи у ПВВбП вибірку $\pi_k^{(1)} = \frac{n}{N} = f$. Але множина осіб (k) тут не є генеральною сукупністю, а насправді є вибірковою сукупністю. Генеральною сукупністю є множина домогосподарств (дг) і саме ця множина досліджується. Множину всіх дг (i) позначимо U_2 та вибірку дг як s_2. Тоді ймовірність включення i-того дг $\pi_i^{(2)}$, до якого належать конкретні m особи $k_1, k_2, ..., k_m$ дорівнює:

$$\pi_i^{(2)} = \mathrm{P}(i \in s_2) = \mathrm{P}(k_1 \cup k_2 \cup ... \cup k_m \in s_1) =$$
$$= \sum_{j=1}^{m} \mathrm{P}(k_j \in s_1) - \sum_{j=1}^{m} \sum_{l=1, l<j}^{m} \mathrm{P}(k_j \cap k_l \in s_1) + ...$$
$$... + (-1)^{m-1} \mathrm{P}(k_1 \cap ... \cap k_m \in s_1).$$

Або можна порахувати ймовірність протилежної події, тобто коли жодні із m осіб, що проживають в i-тому дг не потравлять у вибірку осіб s_1. А після цього знайти шукану ймовірність використовуючи звязок між ймовірністю події та ймовірністю протилежної до неї події.

Якщо $f = 0.1$, тоді:

- для $m = 1$ $\pi_i^{(2)} = 0.1$;
- для $m = 2$ $\pi_i^{(2)} \approx 0.19$;
- для $m = 3$ $\pi_i^{(2)} \approx 0.271$.

1.2 *Відповідь.* Ймовірності включення першого та другого порядків знаходяться за означенням. Коваріаційна матриця індикаторів включення буде мати вигляд

$$\Delta = \frac{1}{36} \times \begin{pmatrix} 5 & -2 & -3 \\ -2 & 8 & -6 \\ -3 & -6 & 9 \end{pmatrix}.$$

1.3 *Вказівка.* Покажіть, що для вибіркових дизайнів із фіксованим розміром вибірки суми значень коваріаційної функції по будь-якому рядку або стовпчику завжди дорівнюють 0.

Ймовірності включення першого порядку можна знайти із діагональних елементів коваріаційної матриці.

Для знаходження ймовірностей включення другого порядку можна скористатися тим, що $\pi_{kl} = \Delta_{kl} + \pi_k\pi_l$ для всіх $k, l \in U$.

1.4 *Вказівка.* Для даного вибіркового дизайну розмір вибірки n_S не є фіксованою величиною, а насправді є випадковою величиною із можливими значенням 2, 3 та 4. Порахуйте ймовірності, що відповідають кожному із цих можливих значень та порахуйте в п.2) $\mathrm{E}n_s$ та $\mathcal{D}n_s$ за означенням. В п.3) знайдіть в підручнику [?] та скористайтесь формулами, що пов'язують $\mathrm{E}n_s$ та $\mathcal{D}n_s$ та ймовірності включення. В п.2) та п.3) ви повинні отримати однакові значення $\mathrm{E}n_s = 3.7$ та $\mathcal{D}n_s = 1.21$.

1.5 *Відповідь.* 1) $\mathrm{E}\widehat{t}_\pi = 3$, $\mathcal{D}\widehat{t}_\pi = 2.69$; 2) $CV(\widehat{t}_\pi) = 0.54$; 3) $\widehat{\mathcal{D}}(\widehat{t}_\pi)(s_1) = 0$, $\widehat{\mathcal{D}}(\widehat{t}_\pi)(s_2) = \widehat{\mathcal{D}}(\widehat{t}_\pi)(s_3) = 10.81$, $\widehat{\mathcal{D}}(\widehat{t}_\pi)(s_4) = 6.57$; 4) $\mathrm{E}\widehat{\mathcal{D}}(\widehat{t}_\pi) = 2.69$.

1.6 *Відповідь.* $\mathrm{E}n_s = 600$ осіб, $\mathcal{D}n_s = 140.801$.

Відповіді та вказівки до розділу 2

2.1 *Відповідь.* $\widehat{t}_\pi = 66\,000$ га; $95\%DI(T) = [64\,231.5; 67786.5]$.

2.2 *Вказівка.* Для генерування рівномірно розподілених на [0,1] чисел скористайтесь таблицею випадкових чисел із Додатку 1.

2.3 *Вказівки.* 1) Шукана ймовірність співпадає із ймовірністю включення першого порядку. 2) Використайте незалежність вибірок. 3) Знайдіть таке n, що ймовірність протилежної до шуканої події буде менше обо дорівнювати 0.5. Відповідь до п.3) 16120 незалежних вибірок.

2.4 *Відповідь.* 0.00123.

2.5 *Відповідь.* $n \geq \frac{z^2_{1-\alpha/2}N^2S^2}{e^2+z^2_{1-\alpha/2}NS^2}$.

2.6 *Вказівка.* 1) Можна скористатись формулою (2.2), де $S^2 = P(1-P)/n$ коли $P = 0.3$; 2) можна виходити із найгіршої ситуації коли вираз $P(1-P)$ набуває свого максимального значення 1/4,

коли $P = 1/2$.

2.7 *Відповідь.* Якщо вважати, що розмір генеральної сукупності вважати нескінченно великим, то необхідний розмір вибірки повинен бути хоча б 3 999 600. Як ви думаєте, чи пов'язано це з тим, що характеристика, яку потрібно виявити, є досить рідкісною?

2.8 *Відповідь.* $n \geq 4900$.

2.9 *Відповідь.* Скористайтесь відповідними формулами для оцінок середнього значення та пропорції. У п.3) подумайте, чи буде працювати ЦГТ.

2.9 *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

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

Тут припускається, що розмір генеральної сукупності такий великий, що можна вважати, що вона нескінченна.

Відповіді та вказівки до розділу 3

3.1 *Відповідь.* 1) $\mathrm{E}n_s = 160$, $\mathcal{D}n_s = 90$; 2) $\pi = \frac{160}{600}$, $\mathcal{D}n_s = 117.3$.

3.2 *Вказівка.* Скористайтесь формулою підрахунку умовної ймовірності.

3.3 *Відповіді.* 1) скористайтесь оцінкою для пропорції при ВБ; 2) $95\%DI(P) = [0.26; 0.30]$; 3) $\widehat{deff}\left(\text{ВБ}, \widehat{P}_{d\pi}\right) = 1.38$. Це означає, що точність відбору Бернуллі гірша, ніж у ПВВбП при такому ж середньому розмірі вибірки.

3.4 *Розв'язок.* 1) Генеральною сукупністю в нас є сукупність ветпунктів, їх в нас $N = 120$. Коефіцієнт варіації оцінки Горвіца–Томпсона $\widehat{t}_\pi$ не повинен перевищувати значення $e = 0.1$. Запишемо умову з якої спочатку знайдемо ймовірність потрапляння елемента у вибірку π. Звертаю увагу, що при визначенні розміру

вибірки ми користуємось теоретичними виразами для дисперсій $\mathcal{D}(\widehat{t}_\pi)$, а не оцінками $\widehat{\mathcal{D}}(\widehat{t}_\pi)$.

$$CV_{\text{ВБ}}(\widehat{t}_\pi) = \frac{\sqrt{\mathcal{D}(\widehat{t}_\pi)}}{T} = \frac{\sqrt{\frac{1-\pi}{\pi}\sum_{k\in U} y_k^2}}{\sum_{k\in U} y_k} \le e.$$

Тоді

$$\frac{1-\pi}{\pi} \le \frac{(e\cdot\sum_{k\in U} y_k)^2}{\sum_{k\in U} y_k^2} = \frac{(0.1\cdot 2400)^2}{50000} = 1.152$$

$$\pi \ge \frac{1}{1.152+1} = 0.4646$$

Отже, мінімальний середній розмір вибірки, що задовольняє нашу умову на кеоефіцінт варіації дорівнює $\mathrm{E}n_s = N\pi = 120\cdot 0.4646 = 55.76$ і може бути не цілим числом.

2)Оскільки при $n = \mathrm{E}n_s$

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

тому обмеження на n отримуємо із нерівності

$$CV_{\text{ВБ}}(\widehat{t}_{\text{альт}}) \approx \frac{\sqrt{N^2(1-\frac{n}{N})\frac{S^2}{n}}}{T} \le e,$$

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

$$= \frac{1}{120-1}\left(50000 - \frac{1}{120}(2400)^2\right) = 16.8;$$

Звідси отримуємо, що

$$n \ge \frac{1}{(\frac{e\cdot T}{N})^2\frac{1}{S^2}+\frac{1}{N}} = \frac{1}{(\frac{0.1\cdot 2400}{120})^2\frac{1}{16.8}+\frac{1}{120}} = 4.057.$$

Отже, середній очікуваний розмір вибірки для альтернативної оцінки $n = \mathrm{E}n_s = 4.057$ і може бути не цілим числом, оскільки ми маємо відбір Бернуллі.

Або, можна скористатись поняттям дизайн-ефект. Дійсно, оскільки дисперсії мало відрізняються $\mathcal{D}_{\text{ВБ}}(\widehat{t}_{\text{альт}}) \approx \mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi)$, тому

$$deff(\text{ВБ}, \widehat{t}_\pi) \approx \frac{\mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi)}{\mathcal{D}_{\text{ВБ}}(\widehat{t}_{\text{альт}})} \Rightarrow \mathcal{D}_{\text{ВБ}}(\widehat{t}_{\text{альт}}) \approx \frac{\mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi)}{deff(\text{ВБ}, \widehat{t}_\pi)}.$$

Тоді

$$CV_{\text{ВБ}}(\widehat{t}_{\text{альт}}) \approx \frac{\sqrt{\frac{1}{deff}\mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi)}}{T} \leq e.$$

Звідси отримуємо, що

$$\mathcal{D}_{\text{ВБ}}(\widehat{t}_\pi) \leq (e \cdot T)^2 deff$$

$$\frac{1-\pi}{\pi} \leq \frac{(e \cdot \sum_{k\in U} y_k)^2 deff}{\sum_{k\in U} y_k^2} = 1.152 \cdot deff$$

Порахуємо яким буде дизайн-ефект в нашому випадку.

$$\overline{Y} = \frac{1}{N}\sum_{k\in U} y_k = \frac{1}{120} \cdot 2400 = 20;$$

$$CV_U = \frac{\sqrt{S^2}}{\overline{Y}} = \frac{\sqrt{16.8}}{20} = 0.2;$$

$$deff \approx 1 + \frac{1}{CV_U^2} = 1 + \frac{\overline{Y}^2}{S^2} = 24.8;$$

$$\pi \geq \frac{1}{29.58} = 0.034,$$

тобто, $n \geq N\pi = 120 \cdot 0.034 = 4.08$, що майже співпадає із значенням, отриманим першим способом. Невелика різниця у значеннях n пов'язана із ефектом заокруглення значень при обчисленнях.

Висновок: альтернативна оцінка в цьому випадку приводить до середнього необхідного розміру вибірки, що в 56/4=14 разів менша, ніж при використанні оцінки Горвіца-Томпсона. Тобто, щоб отримати таку ж точність (яка визначається коефіцієнтом

варіації оцінки) для оцінки Горвіца-Томпсона, яку дає альтернативна оцінка при відборі Бернуллі, ми повинні в середньому обстежити в 14 разів більше елементів. Отже, використання оцінки Г-Т при відборі Бернуллі є дуже неефективним, як і для будь-якого іншого дизайну із нефіксованим розміром вибірки.

Загалом, із виразу для дизайн-ефекту ми бачимо, що чим більш однорідною є генеральна сукупність ($CV_U \approx 0$), тим більшою буде різниця в ефективності використання оцінки Горвіца-Томпсона та альтернативної оцінки. Для неоднорідних сукупностей різниця буде також не на користь оцінки Горвіца-Томпсона, але вже не на стільки значною.

3.5 *Вказівка.* 1) $n \geq 2\,219$. В п.2) потрібно спочатку виразити суму квадратів характеристики y через дисперсію цієї змінної по генеральній сукупності, а саме $\sum_{k\in U} y_k^2 = (N-1)S_y^2 + N\overline{Y}^2$ і підставити це у вираз для коефіцієнта варіації оцінки середнього. Тоді середній необхідний розмір вибірки, що необхідно обстежити при ВБ $En_s \geq 5623.87$. Отже, для отримання потрібної точності, при ВБ потрібно буде обстежити в середньому в 2.5 рази більше елементів, ніж при ПВВбП.

Відповіді та вказівки до розділу 4

4.1 *Вказівка.* Зверніть увагу, що в обох методах систематичні вибірки не будуть неперетинними.

4.2 *Відповідь.* $\widehat{t}_\pi = 600$, $\widehat{\mathcal{D}}_{\text{СВ2}} = 13\,920$. Оцінка дисперсії має досить велике значення - довірчий інтервал навіть заходить у від'ємну піввісь.

4.3 *Вказівка.* Оскільки в задачі дана інформація про всі 10 систематичних вибірки, тому тут можна порахувати реальне значення дисперсії π-оцінки середнього для кожного впорядкування. Зауважимо, що дисперсія оцінки середнього при систематичному відборі може бути підрахована використовуючи формули: $\mathcal{D}(\widehat{\overline{y}}_\pi) = \frac{a(a-1)}{N^2} S_t^2$, $S_t^2 = \frac{n^2}{a-1}\sum_{r=1}^{a}(\overline{y}_{s_r} - \overline{\overline{y}})^2$, $\overline{\overline{y}} = \frac{1}{a}\sum_{r=1}^{a}\overline{y}_{s_r} = \overline{Y}$. Перед використанням вказаних формул, радимо перевірити їх правильність. В п.3) можна використати те, що $SST = (N-1)S^2$ та $SSW = SST - SSB$. 4) Використання зміщеної оцінки при

другому та третьому упорядкуванні буде приводити до побудови консервативних довірчих інтервалів, тобто їх рівень довіри буде насправді більшим, ніж заявлений. А от при першому упорядкуванні заявлений рівень довіри не буде досягатись. Як ви думаєте, яка із цих ситуацій є найменш бажаною?

4.4 *Відповідь.* $\widehat{t}_\pi = 20200$; $\widehat{\mathcal{D}}_{\mathrm{CB10}}\left(\widehat{t}_\pi\right) = 1\,296\,867$; $95\%DI(T) = [17\,624; 22\,776]$. Для побудови довірчого інтервала краще використати квантиль розподілу Стьюдента із 9 ступенями вільності, оскільки випадковість закладено у 10 систематичних вибірок, а для нормального наближення бажано мати хоча б 30 випадкових елементів.

4.5 *Вказівка.* Тут можна скористатись тим, що $S_t^2 = \frac{n}{a-1}SSB$ та $(N-1)S^2 = SST = SSB + SSW$.

Відповіді та вказівки до розділу 5

5.1 *Вказівка.* Знайдіть точний вираз для ймовірності $\mathrm{P}\{k \in os\} = 1 - \mathrm{P}\{k \notin os\}$ та розкладіть цей вираз у ряд Тейлора коли $\frac{1}{N} \to 0$.

5.2 *Вказівка.* 1) $R_1 = \mathrm{P}\{\text{всі елементи os однакові}\} = \frac{1}{N^2}$, $R_3 = \mathrm{P}\{\text{всі елементи os різні}\} = \frac{(N-1)(N-2)}{N^2}$, $R_2 = 1 - R_1 - R_3$. 2) якщо розмір вибірки $n_s = j$, $i = 1, 2, 3$ то умовний вибірковий дизайн можна підрахувати так: $\mathrm{P}(S = s|n_s = i) = \frac{\mathrm{P}(S=s \cap n_s=i)}{R_i} = \frac{1}{R_i}\sum_{os:r(os)=s}\widetilde{p}(os)$, де $\widetilde{p}(os) = \frac{1}{N^3}$. 3) вибірковий дизайн $p(s)$ можна виписати виходячи із отриманого в п. 2) умовного вибіркового дизайну за формулою $p(s) = \mathrm{P}(S = s) = \mathrm{P}(S = s|n_s = i)\mathrm{P}(n_s = i)$. 4) Оскільки оцінка $\overline{y}_1$ є оцінкою Хансена-Гурвіца для середнього $\overline{Y}$, що має властивість незміщеності. Тому її математичне сподівання співпадає із середнім значенням a дисперсія визначається за допомогою формули (5.1.) поділеної на N^2. Щоб порахувати математичне сподівання та дисперсію оцінки $\overline{y}_2$ можна скористатись умовним вибірковим дизайном із п. 2) та формулами $\mathrm{E}(\overline{y}_2) = \mathrm{E}(\mathrm{E}(\overline{y}_2|n_s)) = \mathrm{E}(\overline{y}_2|n_s = 1)R_1 + \mathrm{E}(\overline{y}_2|n_s = 2)R_2 + \mathrm{E}(\overline{y}_2|n_s = 3)R_3 = \overline{Y}$, $\mathcal{D}(\overline{y}_2) = \mathrm{E}(\mathcal{D}(\overline{y}_2|n_s)) + \mathcal{D}(\mathrm{E}(\overline{y}_2|n_s)) = \mathrm{E}(\mathcal{D}(\overline{y}_2|n_s)) = \mathcal{D}(\overline{y}_2|n_s = 1)R_1 + \mathcal{D}(\overline{y}_2|n_s = 2)R_2 + \mathcal{D}(\overline{y}_2|n_s = 3)R_3$.

5.3 *Вказівка.* Якщо припустити, що і генеральна сукупність і вибірка є достатньо великими, щоб добре працювала ЦГТ і мо-

жна було використати квантилі нормального розподілу, тоді $m \geq \frac{z^2_{0.975}}{e^2} P(1-P)$. Виходячи із найгіршої можливої ситуації, коли вираз $P(1-P)$ набуває найбільшого значення 0.25, отримуємо що найменший розмір вибірки, що задовольняє задані рівні точності та надійності, дорівнює $m = 9\,604$.

5.4 *Відповідь.* $p_7 = 5/100 = 0.05$, $p_3 = 0.02$, $p_{56} = 0.01$, $\widehat{t}_{pwr} = 800$, за формулою (5.2) отримаємо $\widehat{\mathcal{D}}(\widehat{t}_{pwr}) = 68333.33$; $\pi_7 = 1 - (1 - p_7)^4 = 0.1855$, $\pi_3 = 0.0776$, $\pi_{56} = 0.0394$, $\widehat{t}_\pi = 529.18$; $\pi_{7\,3} = \pi_7 + \pi_3 - [1-(1-p_7-p3)^4] = 0.0112$, $\pi_{7\,56} = 0.0056$, $\pi_{3\,56} = 0.0023$. Далі залишається лише використати формулу (1.3), з якої отримуємо $\widehat{\mathcal{D}}(\widehat{t}_\pi) = 74494.97$. Кращою є оцінка Хансена-Гурвіца $\widehat{t}_{pwr}$, оскільки її легше підраховувати і вона виявилась точнішою, порівнято із оцінкою Горвіца-Томсона $\widehat{t}_\pi$.

5.5 *Вказівка.* Випадкові числа для відбору можна взяти із Таблиці Д1 в кінці підручника.

Відповіді та вказівки до розділу 6

6.1 *Відповідь.* 1) $\widehat{t}_\pi = 2541,16$; 2) $\widehat{\mathcal{D}}(\widehat{t}_\pi) = 1020956$; 3) $cv(\widehat{t}_\pi) = 0,398$.

6.2 *Відповідь.* 1) $\widehat{\overline{y}}_\pi = 36,97$; 2) $\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi) = 0,0101$; 3) $cv(\widehat{\overline{y}}_\pi) = 0,0027$.

6.3 *Розв'язок.* Згідно зі схемою Сантера, на першому кроці треба впорядкувати $N = 10$ елементів генеральної сукупності у порядку спадання змінної x:

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

На другому кроці 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. Наприклад, нехай $\varepsilon_1 = 0,375$. Тоді $\pi_1 = 4x_1/T_{10} = \frac{4 \cdot 10}{50} = 0,8 > 0,375$, тому елемент 1 потрапляє у вибірку.

Далі, 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. Отже, для $k=2$ маємо $n_2 = 1, \pi_2^{'} = \frac{(4-1)10}{40} = 0,75$. Генеруємо $\varepsilon_2 = 0,624$; $\varepsilon_2 < \pi_2^{'}$, тому елемент 2 потрапляє у вибірку.

Для $k=3$ маємо $n_3 = 2$, $\pi_3^{'} = \frac{(4-2)8}{30} = 0,53$. Генеруємо $\varepsilon_3 = 0,518$; $\varepsilon_3 < \pi_3^{'}$, тому елемент 3 потрапляє у вибірку.

Для $k=4$ маємо $n_4 = 3$, $\pi_4^{'} = \frac{(4-3)6}{22} = 0,273$. Генеруємо $\varepsilon_4 = 0,045$; $\varepsilon_4 < \pi_4^{'}$, тому елемент 4 теж потрапляє у вибірку.

Таким чином, ми отримали вибірку з $n=4$ елементів, тому процедура завершується. При цьому ймовірності включення для відібраних елементів дорівнюють:

$$\pi_1 = \frac{4x_1}{T_{10}} = \frac{4 \cdot 10}{50} = 0,8; \quad \pi_2 = \frac{4x_2}{T_{10}} = \frac{4 \cdot 10}{50} = 0,8;$$

$$\pi_3 = \frac{4x_3}{T_{10}} = \frac{4 \cdot 8}{50} = 0,64; \quad \pi_4 = \frac{4x_4}{T_{10}} = \frac{4 \cdot 6}{50} = 0,48.$$

6.4 *Відповідь.* 1) Результуюча вибірка $\{1,2,4\}$;

2) $\pi_{23} = \frac{9}{25}$, $\pi_{24} = \frac{6}{35}$.

6.5 *Розв'язок.* 1) Ймовірності включення першого порядку залежать від результату вибору елемента на першому кроці: елемент k може бути обраний першим з ймовірністю p_k, тому

$$\pi_k = p_k + (1-p_k)\frac{n-1}{N-1} = p_k\frac{N-n}{N-1} + \frac{n-1}{N-1}, \quad k \in U.$$

Ймовірності включення другого порядку залежать від трьох можливих результатів вибору елемента на першому кроці: вибрано елемент k, або вибрано елемент l, або не вибрано ні k, ні l:

$$\pi_{kl} = p_k\frac{n-1}{N-1} + p_l\frac{n-1}{N-1} + (1-p_k-p_l)\frac{(n-1)(n-2)}{(N-1)(N-2)} =$$

$$= (p_k+p_l)\frac{(n-1)(N-n)}{(N-1)(N-2)} + \frac{(n-1)(n-2)}{(N-1)(N-2)}.$$

2) Скористаємось результатами попереднього пункту:

$$\pi_{kl} = \left(\pi_k \frac{N-1}{N-n} - \frac{n-1}{N-n} + \pi_l \frac{N-1}{N-n} - \frac{n-1}{N-n}\right) \frac{(n-1)(N-n)}{(N-1)(N-2)} +$$

$$+ \frac{(n-1)(n-2)}{(N-1)(N-2)} = \ldots = \frac{(n-1)}{N-2}\left(\pi_k + \pi_l - \frac{n}{N-1}\right).$$

3) Ймовірності включення задовольняють умову Єйтса-Гранді-Сена:

$$\pi_k\pi_l - \pi_{kl} = \left(p_k \frac{N-n}{N-1} + \frac{n-1}{N-1}\right)\left(p_l \frac{N-n}{N-1} + \frac{n-1}{N-1}\right) -$$

$$-(p_k + p_l)\frac{(n-1)(N-n)}{(N-1)(N-2)} - \frac{(n-1)(n-2)}{(N-1)(N-2)} = \ldots =$$

$$= (1 - p_k - p_l)\frac{(n-1)(N-n)}{(N-1)^2(N-2)} + p_k p_l \frac{(N-n)^2}{(N-1)^2} \geq 0.$$

Відповіді та вказівки до розділу 7

7.1 *Розв'язок.* 1)Реальне значення середнього дорівнює 0 (ми знаємо всі значення для характеристики y, тому можемо його порахувати). Справді $\overline{Y} = \frac{1}{N}\sum_{k\in U} y_k = \frac{1}{4}(0+0+1-1) = 0$. Тоді дисперсія генеральної сукупності $S^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \overline{Y})^2 = \frac{1}{3}(0^2+0^2+1^2+(-1)^2) = \frac{2}{3}$. А отже,

$$\mathcal{D}_{\text{ПВВбП}}(\widehat{\overline{y}}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = \left(1 - \frac{2}{4}\right)\frac{2/3}{2} = \frac{1}{6}.$$

2)Для стратифікованого відбору порахуємо відовідні величини по стратам. *Для* U_1: $\overline{Y}_1 = \frac{1}{N_1}\sum_{k\in U_1} y_k = \frac{1}{2}(0+0) = 0$, $S_1^2 = \frac{1}{N_1-1}\sum_{k\in U_1}(y_k - \overline{Y}_1)^2 = \frac{1}{1}(0^2+0^2) = 0$. А отже,

$$\mathcal{D}(\widehat{\overline{y}}_1) = \left(1 - \frac{n_1}{N_1}\right)\frac{S_1^2}{n_1} = \left(1 - \frac{1}{2}\right)\frac{0}{2} = 0.$$

Для U_2: $\overline{Y}_2 = \frac{1}{N_2}\sum_{k\in U_2} y_k = \frac{1}{2}(1+(-1)) = 0$, $S_2^2 = \frac{1}{N_2-1}\sum_{k\in U_2}(y_k-\overline{Y}_2)^2 = \frac{1}{1}((1-0)^2+(-1-0)^2) = 2$. А отже,

$$\mathcal{D}(\widehat{\overline{y}}_2) = \left(1-\frac{n_2}{N_2}\right)\frac{S_2^2}{n_2} = \left(1-\frac{1}{2}\right)\frac{2}{1} = \frac{2}{2} = 1.$$

Тоді,

$$\mathcal{D}_{\text{СТПВВ}}(\widehat{\overline{y}}) = \frac{1}{N^2}\mathcal{D}_{\text{СТПВВ}}(\widehat{t}) = \frac{1}{N^2}\sum_{h=1}^{2} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} =$$

$$= \sum_{h=1}^{2}\frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} = \sum_{h=1}^{2} W_h^2\mathcal{D}(\widehat{\overline{y}}_h) = \left(\frac{2}{4}\right)^2\cdot 0 + \left(\frac{2}{4}\right)^2\cdot 1 = \frac{1}{4}.$$

При цьому, зауважимо, що $\frac{1}{4} > \frac{1}{6}$, а отже стратифікація погіршила якість оцінки середнього.

Висновок: Це означає, що не будь-яка стратифікація призводить до покращення якості оцінок. Потрібно, щоб при стратифікації групи всередині були якомога однорідніші, а між собою групи повинні суттєво відрізнятись. Зокрема, середні значення в них повинні бути різними. А в нас вони були однаковими для обох страт. Друга ж страта була дуже неоднорідна в середині.

7.2 *Розв'язок.* 1)Для підрахунку дисперсії змінної y_k (вага слона) по генеральній сукупності (слони) потрібно спочатку знайти середню вагу слона в цирку:

$$\overline{Y} = \frac{1}{N}\sum_{k\in U} y_k = \frac{1}{N}\left(\sum_{k\in U_1} y_k + \sum_{k\in U_2} y_k\right) =$$

$$= \frac{1}{N}\left(N_1\overline{Y}_1 + N_2\overline{Y}_2\right) = \frac{1}{100}\left(60\times 6 + 40\times 4\right) = \frac{360+160}{100} = 5.2$$

Дисперсію ваги слонів по генеральній сукупності можна порахувати використовуючи основну тотожність дисперсійного аналізу $SST = SSW + SSB$, де $SST = \sum_{k\in U}\left(y_k-\overline{Y}\right)^2$ – загальна варіація, $SSW = \sum_{h=1}^{H}\sum_{k\in U_h}\left(y_k-\overline{Y}_h\right)^2$ – внутрішньогрупова варіація, $SSB = \sum_{h=1}^{H} N_h\left(\overline{Y}_h-\overline{Y}\right)^2$ – міжгрупова варіація. Зокрема, $S^2 = \frac{1}{N-1}SST = \frac{1}{N-1}(SSB+SSW)$.

Порахуємо кожну із цих варіацій окремо:

$$SSB = \sum_{h=1}^{H} N_h \left(\overline{Y}_h - \overline{Y}\right)^2 = N_1(\overline{Y}_1 - \overline{Y})^2 + N_2(\overline{Y}_2 - \overline{Y})^2 =$$
$$= 60(6 - 5.2)^2 + 40(4 - 5.2)^2 = 96;$$

$$SSW = \sum_{h=1}^{H} \sum_{k \in U_h} \left(y_k - \overline{Y}_h\right)^2 = \sum_{h=1}^{H} \frac{N_h - 1}{N_h - 1} \sum_{k \in U_h} \left(y_k - \overline{Y}_h\right)^2 =$$
$$= \sum_{h=1}^{H} (N_h - 1) S_h^2 = 59 \cdot 4 + 39 \cdot 2.25 = 323.75.$$

Отже,

$$S^2 = \frac{1}{N-1} SST = \frac{1}{N-1}(SSB + SSW) = \frac{1}{99}(96 + 323.75) = 4.2399$$

2)ПВВбП, $n = 10$.

$$\mathcal{D}_{\text{ПВВбП}}(\widehat{t}_\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n} = 100^2 \left(1 - \frac{10}{100}\right) \frac{4.2399}{10} = 3815.91$$

3)Порахуємо спочатку розміри вибірок із кожної страти, які потрібно взяти при пропорційному розміщенні 10 елементів:

$$n_1 = n\frac{N_1}{N} = 10\frac{60}{100} = 6, \quad n_2 = n\frac{N_2}{N} = 10\frac{40}{100} = 4$$

Тоді оцінка при стратифікації та пропорційному розміщенні 10 елементів по стратам буде мати дисперсію:

$$\mathcal{D}_{\text{СТПВВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} =$$
$$= 60^2 \left(1 - \frac{6}{60}\right) \frac{4}{6} + 40^2 \left(1 - \frac{4}{40}\right) \frac{2.25}{4} = 2970$$

4) Тепер порахуємо розміри вибірок із кожної страти, які потрібно взяти при нейманівському розміщенні 10 елементів:

$$n_1 = n\frac{S_1N_1}{S_1N_1 + S_2N_2} = 10\frac{\sqrt{4}\cdot 60}{\sqrt{4}\cdot 60 + \sqrt{2.25}\cdot 40} = 6.66$$

$$n_2 = n\frac{S_2N_2}{S_1N_1 + S_2N_2} = 10\frac{\sqrt{2.25}\cdot 40}{\sqrt{4}\cdot 60 + \sqrt{2.25}\cdot 40} = 3.33$$

Заокруглимо ці значення і отримаємо $n_1 = 7$ та $n_2 = 3$. Тоді оцінка при стратифікації та розміщенні Неймана 10 елементів по стратам буде мати дисперсію:

$$\mathcal{D}_{\text{СТПВВ}}(\widehat{t}_\pi) = \sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} =$$
$$= 60^2\left(1 - \frac{7}{60}\right)\frac{4}{6} + 40^2\left(1 - \frac{3}{40}\right)\frac{2.25}{4} = 2927.14$$

Висновок: Виграш у точності при використанні простого стратифікованого відбору та пропорційного розміщення є досить суттєвим, а от використання замість пропорційного розміщення – оптимального розміщення, великого ефекту не дало, а розрахунки при цьому є складнішими та вимагають попередньої інформації про змінну, якої часто немає. Такий ефект зумовлює більшу популярність використання саме пропорційне розміщення при стратифікованому відборі.

7.3 *Відповіді.* 1) $\widehat{\overline{y}}_\pi = 37$; 2) $\mathcal{D}(\widehat{\overline{y}}_\pi) = 0.194$; 3) $n_1 = 60$, $n_2 = 30$, $n_3 = 10$, $\mathcal{D}_{prop}(\widehat{\overline{y}}_\pi) = 0.146$; 4) $C = 1{\cdot}40{+}1{\cdot}20{+}4{\cdot}40 = 220$, $n_1 = 126$, $n_2 = 50$, $n_3 = 11$, $\mathcal{D}_{opt}(\widehat{\overline{y}}_\pi) = 0.082$. Дисперсія зменшилась в більш ніж 2 рази порівняно із початковими результатами.

7.4 *Вказівка.* Вираз $\sum\limits_{h=1}^{H} W_h\overline{y}_h$ співпадає із оцінкою Горвіца-Томсона $\widehat{\overline{y}}_\pi$ при СТПВВ. Якщо ми припустимо, що генеральна сукупність, страти та вибірки в кожній страті є досить великими і ми можемо використати нормальне наближення для розподілу π-оцінки, то потрібну нерівність можна отримати із того, що

$$z\sqrt{\mathcal{D}_{\text{СТПВВ}}(\widehat{\overline{y}}_\pi)} \le a.$$

7.5 *Відповідь.* Якщо $w_1 = 0.2$, то $n \geq 208$; якщо $w_1 = 0.5$, то $n \geq 90$; якщо $w_1 = 0.7$, то $n \geq 84$; якщо $w_1 = 0.8$, то $n \geq 90$. Отже, в околі значення $w_1 = 0.7$ є мінімально можливе значення n, що дозволяє отримати потрібну точність та надійність.

Відповіді та вказівки до розділу 8

8.1 *Відповідь.* Простий випадковий одностадійний кластерний відбір. 1) $\widehat{t}_\pi = 112\,020$; 2) $\widehat{\mathcal{D}}_{\text{ПВОКВ}}(\widehat{t}_\pi) = 117\,800\,190$; 3) $\widehat{\overline{y}}_\pi = 22.404$; $\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi) = 4.712$.

8.2 *Розв'язок.*

1) Такий метод відбору називається простим випадковим кластерним одностадійним відбором, де кластерами є папки із документами клиєнтів. При цьому генеральною сукупністю U у нас є клієнти банку, $N = 39800$, вибіркова сукупність U_I (та, з якої безпосередньо робиться відбір) - це папки (кластери) із погрупованими клієнтами, $N_I = 39\,80$, в кожній папці (кластері) є 10 клієнтів, тобто $N_i = 10$ для всіх $i = \overline{1, N_I}$. Оскільки всі кластери у нас однакові, то середній розмір кластера $\overline{N} = 10$ та $N = \overline{N} N_I$.

2) Частка клієнтів P, що отримали кредит від цього банку – це відношення сумарної кількості клієнтів, яким банк видав кредит, до загальної кількості клієнтів банку. Якщо ми розглянемо змінну

$$z_k = \begin{cases} 1, & \text{якщо клієнт } k \text{ отримав кредит;} \\ 0, & \text{в іншому випадку,} \end{cases}$$

то

$$P = \frac{\sum_{k \in U} z_k}{N} = \frac{\sum_{i \in U_I} \sum_{k \in U_i} z_k}{N} = \frac{\sum_{i \in U_I} A_i}{N} = \frac{\sum_{i \in U_I} A_i}{\overline{N} N_I}.$$

Кластери у вибірку кластерів s_I відбирались ПВВбП із $\pi_{Ii} = \frac{n_I}{N_I}$, де $n_I = 40$, та обстежувались всі клієнти із папки, тому оцінка Горвіца-Томпсона у цьому випадку буде мати вигляд:

$$\widehat{P} = \frac{\frac{N_I}{n_I} \sum_{i \in s_I} A_i}{\overline{N} N_I} = \frac{\sum_{i \in s_I} A_i}{\overline{N} n_I} = \frac{185}{10 \cdot 40} = 0.4625(46.25\%).$$

3) Для того, щоб оцінити дисперсію оцінки частки клієнтів, яким банк надав кредит, використаємо формулу для оцінювання дисперсії оцінки сумарного значення, поділивши її на N^2 (пам'ятаємо, що пропорція є по суті середнім, тобто, дорівнює сумарному значенню, поділеному на N). Тоді з формули (8.7) матимемо

$$\widehat{\mathcal{D}}_{\text{ПВОКВ}}(\widehat{P}) = \frac{N_{\text{I}}^2 \frac{1-f_{\text{I}}}{n_{\text{I}}} \widehat{S}_A^2}{N^2} = \frac{1}{\overline{N}^2} \frac{1-f_{\text{I}}}{n_{\text{I}}} \widehat{S}_A^2,$$

де

$$\widehat{S}_A^2 = \frac{1}{n_I - 1} \left(\sum_{i \in s_I} A_i^2 - \frac{1}{n_I} (\sum_{i \in s_I} A_i)^2 \right) =$$

$$= \frac{1}{39} \left(1236 - \frac{1}{40} (185)^2 \right) = 10.45$$

тоді

$$\widehat{\mathcal{D}}_{\text{ПВОКВ}}(\widehat{P}) = \frac{1}{10^2} \frac{1 - 40/3980}{40} 10.45 = 0.002585 = (0.05)^2,$$

та 95% довірчий інтервал матиме вигляд:

$$95\% DI(P) = [0.4625 \pm 1.96 \cdot 0.05] =$$

$$= [0.3628; 0.5622] = [36.28\%; 56.22\%].$$

8.3 *Відповідь.* Простий випадковий відбір на обох стадіях двостадійного відбору елементів. 1) $\widehat{t}_\pi = 15079.7$; 2) $\widehat{\mathcal{D}}_{\text{ПВДВЕ}}\left(\widehat{t}_\pi\right) = 5173\,334$, $cv(\widehat{t}_\pi) = 0.151$.

8.4 *Відповідь.* 1) $\widehat{y} = \frac{1}{n_I} \sum i \in s_I \overline{y}_i$, де $\overline{y}_i = \frac{1}{n_i} \sum_{k \in s_i} y_k$ – оцінки Горвіца-Томсона середніх витрат всередині i-того кластера (відділення); 2) $\widehat{\mathcal{D}}(\widehat{y}) = \frac{1}{n_I(n_I-1)} \sum_{i \in s_I} (\overline{y}_i - \widehat{y})^2$; 3)$\widehat{y} = 412/025$; $\widehat{\mathcal{D}}(\widehat{y}) = 303.12$.

Відповіді та вказівки до розділу 9

9.1 *Відповідь.* Оцінити різницю D можна за допомогою лінійної оцінки $\widehat{d} = \widehat{t}_{y\pi} - \widehat{t}_{z\pi} = 28\,980$. Оцінка дисперсії $\widehat{d}$ підраховується за формулою $\widehat{\mathcal{D}}(\widehat{d}) = N^2(\frac{1}{n} - \frac{1}{N})[\widehat{S}_y^2 + \widehat{S}_z^2 - 2\widehat{S}_{yz}]$; 95%DI(D)=[22 745; 35 220].

9.2 *Вказівки.* 1) Порівняння ефективності можна робити розглядаючи відношення отриманих двох теоретичних дисперсій. 2) Оцінку Г-Т для параметра$\overline{XY}$ можна записати як $\widehat{\overline{xy}} = \frac{1}{n+n}\left(\sum_{k\in s} x_k + \sum_{k\in s} y_k\right) = \frac{1}{2}(\bar{x} + \bar{y}) = \frac{1}{2n}(\hat{t}_{x\pi} + \hat{t}_{y\pi})$. 3) Тут обчислення робити не обов'язково, хоча і можливо. Певні знання щодо ефективності стратифікації будуть корисними.

9.3 *Розв'язок.* Генеральною сукупністю в нас є сукупність дг, їх в нас $N = 1024$, ПВВбП-вибірка складається із $n = 16$ елементів. Якщо ми розглянемо 2 змінні: z_k – кількість у k-тому дг осіб віком 60+; y_k – кількість у k-тому дг жінок віком 60+, то відсоток жінок серед осіб віком 60+, що проживають в Статиславі є відношенням

$$R = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} = \frac{T_y}{T_z}.$$

Цей параметр ми можемо оцінити підставивши замість сумарних значень T_y, T_z їх оцінки Горвіца-Томпсона $\widehat{t}_y$, $\widehat{t}_z$. Отримаємо

$$\widehat{R} = \frac{\widehat{t}_y}{\widehat{t}_z} = \frac{\frac{N}{n}\sum_{k\in s} y_k}{\frac{N}{n}\sum_{k\in s} z_k} = \frac{\sum_{k\in s} y_k}{\sum_{k\in s} z_k} = \frac{3}{5} = 0.6.$$

а) Для нашого параметра R (відношення) ми вже отримували формули для оцінки дисперсії (9.9) методом лінеаризації Тейлора:

Підрахуємо всі необхідні величини:

$$\overline{z} = \frac{1}{n}\sum_{k\in s} z_k = \frac{1}{16}(1+2+2) = \frac{5}{16} = 0.3125;$$

$$\widehat{S}_y^2 = \frac{1}{n-1}\left(\sum_{k\in s} y_k^2 - \frac{1}{n}(\sum_{k\in s} y_k)^2\right) = \frac{1}{15}(1+1+1-\frac{1}{16}3^2) =$$

$$= \frac{1}{15}(3 - \frac{9}{16}) = 0.1625;$$

$$\widehat{S}_z^2 = \frac{1}{n-1}\left(\sum_{k\in s} z_k^2 - \frac{1}{n}(\sum_{k\in s} z_k)^2\right) = \frac{1}{15}(1+4+4-\frac{1}{16}5^2) =$$
$$= \frac{1}{15}(9-\frac{25}{16}) = 0.4958;$$

$$\widehat{S}_{yz} = \frac{1}{n-1}\left(\sum_{k\in s} y_k z_k - \frac{1}{n}(\sum_{k\in s} y_k)(\sum_{k\in s} z_k)\right) =$$
$$= \frac{1}{15}(1\cdot 2+1\cdot 2+1\cdot 1-\frac{1}{16}3\cdot 5) = \frac{1}{15}(5-\frac{15}{16}) = 0.2708;$$

$$\widehat{\mathcal{D}}(\widehat{R}) = \frac{1}{(0.3125)^2}\frac{1-\frac{16}{1024}}{16}(0.1625+(0.6)^2 0.4958-2\cdot 0.6\cdot 0.2708) =$$
$$= 0.01009764 = (0.100487)^2$$

Якщо виходити з того, що оцінка $\widehat{R}$ є асимптотично нормальною, то шуканий довірчий інтервал матиме вигляд:

$$95\%DI(R) = [0.6 \pm 1.96\cdot 0.100487] = [0.403; 0.797]$$

б)При використанні методу Джек-найф, потрібно порахувати оцінки відношення по вибіркам, що не містять один елемент j, $j = \overline{1,n}$:

$$\widehat{R}_{(j)} = \frac{\sum_{k\neq j} y_k}{\sum_{k\neq j} z_k}.$$

Для 13 дг, де не має осіб віком 60+ $z_j = y_j = 0$ та $\sum_{k\neq j} y_k = \sum_{k\in s} y_k = 3$, $\sum_{k\neq j} z_k = \sum_{k\in s} z_k = 5$, $\widehat{R}_{(j)} = \frac{3}{5} = 0.6$.

Для тих 2 дг, де проживають по 2 осіб віком 60+, одна з яких жінка $z_j = 2$, $y_j = 1$ та $\sum_{k\neq j} y_k = \sum_{k\in s} y_k - 1 = 3-1 = 2$, $\sum_{k\neq j} z_k = \sum_{k\in s} z_k - 2 = 5-2 = 3$, $\widehat{R}_{(j)} = \frac{2}{3} = 0.66$.

Для 1 дг, де проживає одна жінка віком 60+, $z_j = 1$, $y_j = 1$ та $\sum_{k\neq j} y_k = \sum_{k\in s} y_k - 1 = 3-1 = 2$, $\sum_{k\neq j} z_k = \sum_{k\in s} z_k - 2 = 5-1 = 4$, $\widehat{R}_{(j)} = \frac{2}{4} = 0.5$.

Використаємо другий варіант JK оцінки для оцінки дисперсії:

$$\begin{aligned}\widehat{\mathcal{D}}_{JK2}(\widehat{R}) &= \frac{n-1}{n}\sum_{j=1}^{n}(\widehat{R}_{(j)}-\widehat{R})^2 = \\ &= \frac{15}{16}(13(0.6-0.6)^2+2(0.66-0.6)^2+(0.5-0.6)^2) = \\ &= \frac{15}{16}(2(0.06)^2+(0.1)^2) = 0.016125.\end{aligned}$$

Якщо ще зробити поправку на скінченність генеральної сукупності, то отримаємо

$$\widehat{\mathcal{D}}'_{JK2}(\widehat{R}) = (1-\frac{16}{1024})0.016125 = 0.01587 = (0.126)^2$$

Якщо виходити з того, що оцінка $\widehat{R}$ є асимптотично нормальною, то шуканий довірчий інтервал матиме вигляд:

$$95\%DI(R) = [0.6 \pm 1.96 \cdot 0.126] = [0.35; 0.84]$$

в) Для оцінювання дисперсії методом Бут-стреп створимо штучну генеральну сукупність, кожен елемент якої буде утворений із N/n=1024/16=64 копій значень та y із вибірки:

- значення z для вибірки складються із 2-х '2', однієї '1' та всі інші '0', отже наша штучна сукупність буде мати по характеристиці z $2 \cdot 64 = 128$ – '2', $1 \cdot 64 = 64$ – '1', всі інші $13 \cdot 64 = 832$ – '0'.

- значення y для вибірки складається із 3-х '1' та всі інші '0', отже наша штучна сукупність буде мати по характеристиці y $3 \cdot 64 = 192$ – '1', та всі інші $13 \cdot 64 = 832$ – '0'.

Тоді, якщо ми виберемо із нашої штучної сукупності U^* ПВВбП вибірку $s_1 = \{659, 151, 137, 90, 772, 5, 315, 144, 564, 302, 683, 483, 106, 696, 14, 978\}$ (функція `sample(1:1024,16)`) та порахуємо по елементам із цієї вибірки значення нашого параметра, то отримаємо $\widehat{R}_1 = 0.6666667$.

Цю процедуру виконуємо 1000 разів для різних бут-стреп вибірок, беручи до уваги, що для деяких випадків ми можемо отримувати ділення на 0. Ми отримуємо вектор, що складається із 1000 значень оцінки $\widehat{R}_r$, для якого обчислюємо значення $\widehat{R}^* = 0.617371$ та $\widehat{\mathcal{D}}_{BS}(\widehat{R}) = 0.01849171$.

Якщо виходити з того, що оцінка $\widehat{R}$ є асимптотично нормальною, то шуканий довірчий інтервал матиме вигляд:

$$95\%DI(R) = [0.6 \pm 1.96 \cdot \sqrt{0.01849171} = [0.33; 0.86]$$

Для отримання ДІ із розподілу $\widehat{R}$ можна скористатись функцією `quantile(..., probs = c(0.025,0.975))`. В результаті застосування цієї функції отримуємо $95\%DI(R) = [0.5; 1]$.

9.4 *Відповіді.* 1) $\widehat{d}$=3 345.54, $95\%DI(D) = [2695.19; 3995.88]$. Позитивний ефект від рекламної кампанії безумовно є, але за перший місяць кошти, що були витрачені на неї, не були компенсовані.

2) $\widehat{R} = 1.22 > 1$, тому і цей показник вказує на позитивний ефект рекламної кампанії. Довірчий інтервал $95\%DI(R) = [1.06; 3.51]$ говорить, про те, що реальне збільшення продажів може коливатись від 6% до 351%. Це досить широкий діапазон. Для уточнення ефекту потрібні додаткові дослідження.

9.5 *Вказівка.* Рівень успішності для студентів чи студенток є, в цьому випадку, параметром відношення R, що не є лінійним. Тому цей параметр має зміщення. Щоб визначити приблизно яким є це зміщення, можна використати такий розклад:

$$\widehat{R} = \frac{\overline{y}}{\overline{z}} = \frac{\overline{Y}\frac{\overline{y}}{\overline{Y}}}{\overline{Z}\frac{\overline{z}}{\overline{Z}}} = R\frac{1 - \frac{\overline{y}-\overline{Y}}{\overline{Y}}}{1 - \frac{\overline{z}-\overline{Z}}{\overline{Z}}} = R\frac{1-\delta_y}{1-\delta_z} = R(1-\delta_y)(1-\delta_z+\delta_z^2-...).$$

Якщо в цьому розкладі взяти перші декілька членів, то можна показати, що при ПВВбП зміщення $B(R) \approx R(1-f)\frac{1}{n}[\frac{S_z^2}{\overline{Z}^2} - \frac{S_{yz}}{\overline{Y}\overline{Z}}]$. Залишається лише порахувати якого вигляду набудуть всі складові цього виразу, коли ми маємо справу із реальною успішністю, і обчислити в цьому випадку його значення. Отримаємо, що $B(R) \approx 0$.

9.6 *Вказівка.* У п.2 запропонована оцінка є по суті оцінкою відношення. Дісно, якщо розглянути змінну z_k, що набуває значення 1, якщо k-те дг має дитину до 1 року та 0, якщо не має, тоді $n^* = \sum_{k\in s} z_k$.

Відповіді та вказівки до розділу 10

10.1 *Вказівка.* Оцінка за відношенням має вигляд $\widehat{t}_{yR} = T_x\widehat{B}$, де $\widehat{B} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}} = \frac{\sum_{k\in s} y_k}{\sum_{k\in s} x_k}$ – є оцінкою відношення. Тому формули для дисперсії та її оцінки при ПВВбП можна досить легко отримати беручи до уваги те, що $\mathcal{D}(\widehat{t}_{yR}) = \mathcal{D}(T_x\widehat{B}) = T_x^2\mathcal{D}(\widehat{B})$.

10.2 *Відповідь.* $\widehat{t}_{yR} = 136\,885.9$; $\widehat{\mathcal{D}}(\widehat{t}_{yR}) = 12\,544\,502$.

10.3 *Розв'язок.* Генеральною сукупністю в нас є сукупність дг, їх в нас $N = 1024$, ПВВбП-вибірка складається із $n = 20$ елементів. Розглянемо 2 змінні: y_k – кількість у k-тому дг осіб віком 60+; $x_k = 1$ – якщо у k-тому дг є особи віком 60+, та $x_k = 0$, якщо такі особи в дг не проживають.

Результати нашого обстеження подауj у вигляді таблиці:

k	1	2	3	4	...	20
y_k	2	2	1	0	...	0
x_k	1	1	1	0	...	0

Крім того, нам відомо додатково, що $T_x = \sum_{k\in U} x_k = 300$.

1) Оцінимо шукану кількість за допомогою оцінки Горвіца-Томпсона: $\widehat{t}_{y\pi} = N\overline{y} = 1024 \cdot 0.25 = 256$.

$$\widehat{S}_y^2 = \frac{1}{n-1}\left(\sum_{k\in s}(y_k - \overline{y})^2\right) =$$

$$= \frac{1}{19}\left(2(2-0.25)^2 + (1-0.25)^2 + 17(0-0.25)^2\right) = 0.4079$$

$$\widehat{\mathcal{D}}(\widehat{t}_{y\pi}) = N^2(1-f)\frac{\widehat{S}_y^2}{n} = 1024^2(1-20/1024)\frac{0.4079}{20} = 20967.75$$

2) Використаємо оцінку за різницею: $\widehat{t}_{yD} = T_x + (\widehat{t}_{y\pi} - \widehat{t}_{x\pi}) = 300 + (256 - 153.6) = 402.4$ оскільки $\widehat{t}_{x\pi} = N\overline{x} = 1024 \cdot 0.15 = 153.6$.

Для підрахунку оцінки дисперсії оцінки сумарного за різницею, спочатку підрахуємо всі необхідні для цього величини:

$$\widehat{S}_x^2 = \frac{1}{n-1}\left(\sum_{k\in s}(x_k - \overline{x})^2\right) = \\ = \frac{1}{19}\left(3(1-0.15)^2 + 17(0-0.15)^2\right) = 0.1342$$

$$\widehat{S}_{xy} = \frac{1}{n-1}\left(\sum_{k\in s} x_k y_k - \frac{1}{n}\left(\sum_{k\in s} x_k\right)\left(\sum_{k\in s} y_k\right)\right) = \\ = \frac{1}{19}\left(2\cdot 1 + 2\cdot 1 + 1\cdot 1 + 17\cdot 0\cdot 0 - \frac{5\cdot 3}{20}\right) = 0.2236$$

Отже,

$$\widehat{\mathcal{D}}(\widehat{t}_{yD}) = N^2(1-f)\frac{1}{n}\left(\widehat{S}_y^2 + \widehat{S}_y^2 - 2\widehat{S}_{xy}\right) = \\ = 1024^2(1-20/1024)\frac{0.40789 + 0.1342 - 2\cdot 0.2236}{20} = 4869.92$$

Висновок: використовуючи додаткову інформацію за допомогою оцінювання за різницею ми отримали ефект зменшення дисперсії у більш ніж 4 рази порівняно із π-оцінкою. Дійсно

$$\frac{\widehat{\mathcal{D}}(\widehat{t}_{y\pi})}{\widehat{\mathcal{D}}(\widehat{t}_{yD})} = \frac{20967.75}{4869.92} = 4.3.$$

Таке суттєве покращення якості оцінки зумовлене тим, що коефіцієнт кореляції між змінними досить великий. Зокрема $\widehat{r} = \frac{\widehat{S}_{xy}}{\widehat{S}_x\widehat{S}_y} = 0.956$ що і дало такий суттєвий ефект.

3) Порахуємо спочатку оцінку відношення (не плутати із оцінкою *по* відношенню!)

$$\widehat{B} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}} = \frac{256}{153.6} = 1.67 \quad \text{та} \quad \widehat{t}_{yR} = \widehat{B}\cdot T_x = 1.67\cdot 300 = 500.$$

Тоді

$$\widehat{\mathcal{D}}(\widehat{t}_{yR}) = \frac{T_x^2}{\widehat{t}_{x\pi}^2} N^2(1-f)\frac{1}{n}\left(\widehat{S}_y^2 + \widehat{B}^2\widehat{S}_x^2 - 2\widehat{B}\widehat{S}_{xy}\right) =$$
$$= \frac{(300)^2 1024^2}{(153.6)^2}(1-\frac{20}{1024})\frac{0.4079 + (1.67)^2 0.1342 - 2\cdot 1.67\cdot 0.2236}{20} =$$
$$= 5379.5$$

Висновок: використовуючи додаткову інформацію за допомогою оцінювання за відношенням ми отримали ефект зменшення дисперсії, але у менше ніж 4 рази. Дійсно

$$\frac{\widehat{\mathcal{D}}(\widehat{t}_{y\pi})}{\widehat{\mathcal{D}}(\widehat{t}_{yD})} = \frac{20967.75}{5379.5} = 3.89.$$

Порівняно із оцінкою Горвіца-Томпсона ми отремали кращий результат, але якщо порівняти із оцінкою за різницею – то результат погіршився. Це може бути пов'язним з тим, що модель, що відповідає оцінці по відношенню не зовсім адекватно (гірше) відображає залежність змінних y та x.

4) При використанні оцінки за лінійною регресією спочатку треба порахувати коефіцієнти $\widehat{B_1} = \frac{\widehat{S}_{xy}}{\widehat{S}_x^2} = \frac{0.2236}{0.1342} = 1.667$, $\widehat{B}_0 = \overline{y} - \widehat{B}_1\overline{x} = \frac{1}{n}(\widehat{t}_{y\pi} - \widehat{B}_1\widehat{t}_{x\pi}) = \frac{1}{20}(256 - 1.667\cdot 153.6) = 0.076$. Тоді, оцінка сумарного за лінійною регресією буде дорівнювати

$$\widehat{t}_{yLR} = \widehat{t}_{y\pi} + \widehat{B}_1(T_x - \widehat{t}_{x\pi}) =$$
$$= 256 + 1.667(300 - 153.6) = 500$$

Для підрахунку оцінки дисперсії нам потрібна буде величина

$$\widehat{S}_e^2 = \frac{1}{n-1}\left(\sum_{k\in s}(y_k - \widehat{B}_0 - \widehat{B}_1 x_k)^2\right) =$$
$$= \frac{1}{19}\left(2(2 - 0.076 - 1.667\cdot 1)^2 + (1 - 0.076 - 1.667\cdot 1)^2 + \right.$$
$$\left. + 17(0 - 0.076 - 1.667\cdot 0)^2\right) = 0.0412.$$

Отже,

$$\widehat{\mathcal{D}}(\widehat{t}_{yLR}) = N^2(1-f)\frac{1}{n}\widehat{S}_e^2 =$$
$$= 1024^2(1-20/1024)\frac{0.0412}{20} = 2117.878$$

Висновок: використовуючи додаткову інформацію за допомогою оцінювання за лінійною регресією ми отримали ефект зменшення дисперсії у майже 10 разів. Дійсно

$$\frac{\widehat{\mathcal{D}}(\widehat{t}_{y\pi})}{\widehat{\mathcal{D}}(\widehat{t}_{yLR})} = \frac{20967.75}{2117.878} = 9.9.$$

10.4 *Вказівка.* Дивіться розв'язок попередньої задачі 10.3

10.5 *Відповідь.* 1) Тут можна використати оцінку Горвіца-Томсона: $\widehat{\overline{y}}_\pi = 1.8$, $\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi) = 0.2255$, $95\%DI(\overline{Y}) = [0.87; 2.73]$. Точність отриманої оцінки не найкраща. Причина у надто малому розмірі вибірки. Також рівень надійності побудованого довірчого інтервалу може відрізнятись від заявленого, оскільки при такому невеликому розмірі генеральної сукупності та розмірі вибірки нормальна апроксимація працює погано.

2) Якщо розглянути додаткову змінну x, що набуває значення 1, якщо дитина має хоча б один зуб із карієсом та 0 в іншому випадку, тоді можна скористатись оцінкою за відношенням. $\widehat{\overline{y}}_R = \overline{X}\frac{\overline{y}}{\overline{x}} = 2.25$; $\widehat{\mathcal{D}}(\widehat{\overline{y}}_R) = 0.1622$. Ефект у точності можна проаналізувати якщо розглянути відношення

$$\frac{\widehat{\mathcal{D}}(\widehat{\overline{y}}_R)}{\widehat{\mathcal{D}}(\widehat{\overline{y}}_\pi)} = \frac{0.1622}{0.2255} = (0.85)^2.$$

Отже, залучення додаткової інформації через оцінку по відношенню, що в цьому випадку співпадає із пост-страфікаційною оцінкою, дало зменшення ширини довірчого інтервалу на 15%.

3) З одніє сторони, залучення додаткової інформації дало ефект покращення точності оцінювання, але залучення додаткового спеціаліста вимагає збільшення фінансових витрат. І хоча у другого дантиста обстеження кожної дитини проходить швидше,

оскільки йому не потрібно рахувати зуби із карієсом, але він повинен обстежити в 10 разів більше дітей. Тому ефект в покращенні точності оцінки при залученні додаткової інформації треба ще порівнювати із додатковими фінансовими затратами, що виникають при цьому.

Додаток 1. Таблиця випадкових чисел

Числа, наведені у таблиці Д 1, можна використовувати для отримання ймовірнісної вибірки.

Таблиця Д 1. **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 Д 2 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 Д 2 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

Список основних позначень

$\mathrm{E}(\widehat{\theta})$ – 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. *Василик, О. І.* Лекції з теорії і методів вибіркових обстежень / О. І. Василик, Т. О. Яковенко. – К. : ВПЦ «Київський ун-т», 2010.

2. Вибіркове спостереження. Термінологічний словник / О. О. Васєчко, О. І. Черняк, Є. М. Жуйкова та ін. – К. : ІВЦ ДКС України, 2004.

3. *Зінченко, Н. М.* Аналітичні моделі та методи соціології / Н. М. Зінченко, А. Я. Оленко. – К. : ВПЦ «Київський ун-т», 2000.

4. *Карташов, М. В.* Імовірність, процеси, статистика / М. В. Карташов. – К. : ВПЦ «Київський ун-т», 2007.

5. *Майборода, Р. Є.* Регресія: лінійні моделі / Р. Є. Майборода. – К. : ВПЦ «Київський ун-т», 2007.

6. *Пархоменко, В. М.* Методи вибіркових обстежень / В. М. Пархоменко. – К. : ТВіМС, 2001.

7. *Саріогло, В. Г.* Проблеми статистичного зважування вибіркових даних / В. Г. Саріогло. – К. : ІВЦ Держкомстату України, 2005.

8. *Черняк, О. І.* Техніка вибіркових досліджень / О. І. Черняк. – К. : МІВВЦ, 2001.

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. *Bulmer, M.* Life on an island: A simulated population to support student projects in statistics / M. Bulmer, J. K. Haladyn// Technology Innovations in Statistics Education, 2011. – Vol. 5, No. 1.

13. *Brewer, K. R. W.* Sampling with Unequal Probabilities / K. R. W. Brewer, M. Hanif. – Springer-Verlag, 1983.

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. *Schwarz, C. J.* StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling / C. J. Schwarz// Journal of Statistics Education, 1997. – Vol. 5, No. 2.

22. *Sunter, A. B.* Response burden, sample rotation and classification renewal in economic surveys / A. B. Sunter // International Statistical Review, 1977. – Vol. 45.

23. *Sunter, A. B.* List sequential sampling with equal or unequal probabilities without replacement / A. B. Sunter // Applied Statistics, 1977. – Vol. 26, No. 3.

24. *Thompson, S.-K.* Sampling, 3rd Edition / S.-K. Thompson. – Wiley, 2012.