**Міністерство освіти і науки України**
**Дніпровський національний університет**
**імені Олеся Гончара**

**Я. С. Бондаренко**

# ПОСІБНИК ДО ВИВЧЕННЯ ДИСЦИПЛІНИ “ВИБІРКОВІ ОБСТЕЖЕННЯ”

**Дніпро**
**2018**

**УДК 519.253 (075.8)**
**Б81**

Рецензенти: канд. фіз.-мат. наук, доц. В.М. Трактинська,
канд. фіз.-мат. наук, доц. Н.І. Послайко

Б81 Бондаренко, Я. С. Посібник до вивчення дисципліни “Вибіркові обстеження” [Текст] / Я.С. Бондаренко. – Д.: Ліра, 2018. – 24 с.

Викладені основні поняття і факти щодо стратифікованого відбору. Теоретичні положення проілюстровані прикладами.

Для студентів механіко-математичного факультету ДНУ спеціальності “Статистика”.

*Рекомендовано до друку вченою радою*
*механіко-математичного факультету*
*Дніпровського національного університету імені Олеся Гончара*
*протокол №2 від 23.10.2018 року*

Навчальне видання

Яна Сергіївна Бондаренко

**Посібник до вивчення дисципліни**
**“Вибіркові обстеження”**

Друкується за авторською редакцією

---

Підписано до друку 30.10.18. Формат 60×84/16. Папір друкарський. Друк плоский. Ум. друк. арк. 1,4. Тираж 50 пр. Зам. № 310.

---

Друкарня «Ліра», вул. Наукова, 5, м. Дніпро, 49107.
Свідоцтво про внесення до Державного реєстру серія ДК №6042 від 26.02.2018 р.

# ВСТУП

Наші знання, судження, вчинки ґрунтуються на вибіркових даних. Нам доступний для вивчення лише фрагмент загальної картини, який повинен розширити наші знання. Важливо знати як правильно здобути вибірку і як зробити за її даними обґрунтовані висновки. Ця проблема не відіграє особливої ролі, якщо популяція, з якої проводиться відбір, є однорідною. Проте, коли популяція неоднорідна, спосіб здобуття вибірки набуває визначного значення, а вивчення методів, які дозволяють отримати достовірні результати, безумовно стає актуальним.

Досить часто на практиці при проведенні вибіркового обстеження доступна деяка додаткова інформація щодо популяції, яка досліджується. Наприклад, відомо, що середнє значення величини, яка цікавить дослідника, суттєво відрізняється для деяких підпопуляцій. У такому випадку можливо дістати більш точні оцінки величини, що характеризує популяцію, за допомогою використання *стратифікованого відбору*.

При стратифікованому відборі популяція, що містить $N$ одиниць, спочатку поділяється на підпопуляції, що складаються відповідно з $N_1,\ N_2,\ \dots,\ N_L$ одиниць. Ці підпопуляції не містять спільних одиниць так, що

$$N_1 + N_2 + \cdots + N_L = N.$$

Такі підпопуляції називають *стратами.* Для того, щоб можна було повністю скористатися вигодами від стратифікації, значення $N_1, N_2, \dots, N_L$ повинні бути відомі. За визначених страт, вибірка добувається з кожної страти, при цьому відбір у різних стратах відбувається незалежно. Обсяги вибірок всередині страт позначаються через $n_1,\ n_2,\ \dots,\ n_L$ відповідно.

Якщо у кожній страті здобувають просту випадкову вибірку, то спосіб відбору називається *стратифікованим випадковим відбором*.

Стратифікація є досить розповсюдженим методом. Це зумовлено багатьма причинами; наведемо основні з них.

1. Якщо бажано одержати дані про деякі підпопуляції популяції з певною точністю, то кожну таку підпопуляцію рекомендується розглядати на правах самостійної популяції.

2. Застосування стратифікації може бути продиктоване організаційними міркуваннями, наприклад, агентство, що здійснює вибіркове обстеження, може мати відділи, кожен з яких забезпечує проведення вибіркового обстеження деякої частини популяції.

3. Проблеми, пов’язані з відбором у різних частинах популяції, суттєво різняться. При вибіркових обстеженнях осіб, що перебувають у таких закладах, як готелі, лікарні, в'язниці, часто виділяють в окрему страту на відміну від осіб, що мешкають у звичайних будинках, оскільки до відбору в цих двох випадках необхідний різний підхід. При вибірковому обстеженні, розпочатому з метою вивчення ділової активності, ми можемо скласти список великих фірм, відокремити їх в окрему страту, а для дрібних фірм застосувати один з видів територіального відбору.

4. Стратифікація дає виграш у точності при оцінюванні характеристик усієї популяції. Іноді неоднорідну популяцію вдається розділити на підпопуляції, кожна з яких внутрішньо однорідна. Якщо кожна страта однорідна в тому сенсі, що результати вимірювань в ній досить мало змінюються від одиниці до одиниці, то можна одержати точну оцінку середнього значення для будь-якої страти за невеликою вибіркою з цієї страти, а потім ці оцінки можна об’єднати в одну оцінку для всієї популяції.

При стратифікованому відборі розглядаються властивості оцінок, знайдених за стратифікованими вибірками, та умови визначення найкращих обсягів вибірок за стратами для отримання найбільшої точності. Вважається, що самі страти вже створені.

Для повного використання можливостей техніки стратифікованого випадкового обстеження необхідно розв'язати деякі практичні питання побудови стратифікованої вибірки. Перше питання полягає у поділі всієї популяції на страти, друге – у виборі методу здобуття випадкової вибірки і методу оцінювання для кожної страти.

Введемо наступні позначення. Великі літери відносяться до характеристик популяції, малі – до характеристик вибірки. Індекси $h$ та $i$ відповідають номеру страти та номеру одиниці в страті. Нехай $N$ – число одиниць популяції; $n$ – число одиниць у вибірці; частка відбору з популяції

$$f = \frac{n}{N};$$

$N_h$ – загальне число одиниць в страті $h$; $n_h$ – число одиниць у вибірці зі страти $h$; частка відбору в страті $h$

$$f_h = \frac{n_h}{N_h};$$

вага страти $h$

$$W_h = \frac{N_h}{N};$$

середнє значення змінної $y$ в страті $h$ популяції

$$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi};$$

дисперсія змінної $y$ в страті $h$ популяції

$$S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2;$$

середнє значення змінної $y$ в популяції

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_h;$$

дисперсія змінної $y$ в популяції

$$S^2 = \frac{1}{N - 1}\sum_{h=1}^{L}\sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2;$$

вибіркове середнє значення змінної $y$ в страті $h$

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi};$$

вибіркова дисперсія змінної $y$ в страті $h$

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

# 1. Стратифікований випадковий відбір

## 1.1. Оцінювання середнього значення

При простому випадковому відборі вибіркове середнє є незміщеною оцінкою середнього значення популяції. При стратифікованому відборі за оцінку середнього значення популяції розглядається оцінка:

$$\bar{y}_{st} = \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2 + \cdots + \frac{N_L}{N}\bar{y}_L = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{y}_h = \sum_{h=1}^{L} W_h\bar{y}_h,$$

де $N_1 + N_2 + \cdots + N_L = N$.

Оцінка $\bar{y}_{st}$, взагалі кажучи, не збігається з вибірковим середнім

$$\bar{y} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_2 + \cdots + \frac{n_L}{n}\bar{y}_L = \frac{1}{n}\sum_{h=1}^{L} n_h\bar{y}_h.$$

Відмінність полягає у тому, що в оцінці $\bar{y}_{st}$ оцінкам $\bar{y}_h$, здобутим за окремими стратами, надаються ваги $W_h$. Очевидно, що оцінка $\bar{y}$ збігається з оцінкою $\bar{y}_{st}$ за умов, що для кожної страти

$$\frac{n_h}{n} = \frac{N_h}{N}$$

або

$$\frac{n_h}{N_h} = \frac{n}{N}.$$

Останнє означає, що частка відбору однакова для всіх страт. Така стратифікація називається *стратифікацією з пропорційним розміщенням*. Вона забезпечує рівнозважену вибірку.

Основні властивості оцінки $\bar{y}_{st}$ викладено у наступних теоремах.

*Зауваження.* Перші дві теореми відносяться до стратифікованого відбору вцілому, а не тільки до стратифікованого випадкового відбору; іншими словами, вибірка з кожної страти не обов'язково повинна бути простою випадковою вибіркою.

**Теорема 1.1.1**. Якщо для кожної страти вибіркове середнє $\bar{y}_h$ є незміщеною оцінкою $\bar{Y}_h$, то $\bar{y}_{st}$ є незміщеною оцінкою середнього значення популяції $\bar{Y}$.

Доведення. Математичне сподівання оцінки $\bar{y}_{st}$ дорівнює

$$M\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h M\bar{y}_h = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_h,$$

оскільки оцінки за окремими стратами незміщені. Середнє значення змінної $y$ в популяції можна записати у вигляді:

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi} = \frac{N_1}{N}\bar{Y}_1 + \frac{N_2}{N}\bar{Y}_2 + \cdots + \frac{N_L}{N}\bar{Y}_L = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_h$$

отже, $\bar{y}_{st}$ – незміщена оцінка середнього значення популяції $\bar{Y}$.

**Наслідок**. Оскільки при простому випадковому відборі всередині страт вибіркові середні $\bar{y}_h$ є незміщеними оцінками середніх значень $\bar{Y}_h$, то при *стратифікованому випадковому відборі* оцінка $\bar{y}_{st}$ є незміщеною оцінкою середнього значення $\bar{Y}$ популяції.

**Теорема 1.1.2**. При стратифікованому відборі дисперсія оцінки $\bar{y}_{st}$ середнього значення популяції $\bar{Y}$ має вигляд:

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 D\bar{y}_h = \sum_{h=1}^{L} W_h^2 D\bar{y}_h,$$

де $D\bar{y}_h = M(\bar{y}_h - M\bar{y}_h)^2 = M(\bar{y}_h - \bar{Y}_h)^2$.

Доведення. Розглянемо різницю:

$$\bar{y}_{st} - \bar{Y} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h - \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_h = \frac{1}{N}\sum_{h=1}^{L} N_h(\bar{y}_h - \bar{Y}_h).$$

Відмітимо, що помилка оцінки, $\bar{y}_{st} - \bar{Y}$, подана як зважене середнє помилок, $\bar{y}_h - \bar{Y}_h$, здобутих при оцінюванні за окремими стратами. Отже,

$$(\bar{y}_{st} - \bar{Y})^2 = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2(\bar{y}_h - \bar{Y}_h)^2 + \frac{2}{N^2}\sum_{h,j\ h\neq j} N_h N_j(\bar{y}_h - \bar{Y}_h)\big(\bar{y}_j - \bar{Y}_j\big),$$

причому в останньому доданку підсумовування проводиться за всіма парами страт.

Математичне сподівання квадрату помилки оцінки дорівнює

$$M(\bar{y}_{st} - \bar{Y})^2 = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 M(\bar{y}_h - \bar{Y}_h)^2 + \frac{2}{N^2}\sum_{h,j\ h\neq j} N_h N_j M(\bar{y}_h - \bar{Y}_h)\big(\bar{y}_j - \bar{Y}_j\big).$$

Оскільки відбір за стратами $h$ та $j$ відбувається незалежно і $\bar{y}_j$ є незміщенною оцінкою для $\bar{Y}_j$, то $M(\bar{y}_j - \bar{Y}_j) = 0$. Тому усі доданки з індексами, що не збігаються, дорівнюють нулеві. Отже,

$$D\bar{y}_{st} = M(\bar{y}_{st} - \bar{Y})^2 = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 M(\bar{y}_h - \bar{Y}_h)^2 = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 D\bar{y}_h = \sum_{h=1}^{L} W_h^2 D\bar{y}_h.$$

Важлива особливість цього результату полягає в тому, що дисперсія $D\bar{y}_{st}$ залежить тільки від дисперсій $D\bar{y}_h$ вибіркових середніх значень окремих страт.

**Теорема 1.1.3.** При *стратифікованому випадковому відборі* дисперсія оцінки $\bar{y}_{st}$ середнього значення $\bar{Y}$ популяції має вигляд:

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2(1 - f_h)\frac{S_h^2}{n_h}.$$

Доведення. Оскільки $\bar{y}_h$ є незміщеною оцінкою $\bar{Y}_h$, то можна застосувати теорему 1.1.2. Відомо, що дисперсія вибіркового середнього для простої випадкової вибірки в страті $h$ дорівнює:

$$D\bar{y}_h = \frac{S_h^2}{n_h}\frac{N_h - n_h}{N_h}.$$

Підставимо $D\bar{y}_h$:

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 D\bar{y}_h = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2(1 - f_h)\frac{S_h^2}{n_h}.$$

Деякі частинні випадки цієї формули наведені у наслідках 1, 2, 3.

**Наслідок 1.** Якщо частки відбору $f_h$ в усіх стратах малі, то

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}.$$

Доведення. Дійсно,

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}.$$

Оскільки частки відбору $f_h$ малі, то

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}.$$

**Наслідок 2.** Для стратифікованої вибірки з пропорційним розміщенням дисперсія набуває вигляду

$$D\bar{y}_{st} = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2.$$

Доведення. Дійсно,

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \frac{1}{N^2}\sum_{h=1}^{L} N_h\left(N_h - \frac{nN_h}{N}\right)\frac{NS_h^2}{nN_h},$$

$$D\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\left(1 - \frac{n}{N}\right)\frac{S_h^2}{n} = \sum_{h=1}^{L}\frac{N_h}{N}\left(\frac{N-n}{N}\right)\frac{S_h^2}{n} = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2.$$

**Наслідок 3.** Для стратифікованої вибірки з пропорційним розміщенням та однаковою дисперсією $S_h^2$ в усіх стратах

$$D\bar{y}_{st} = \frac{S_h^2}{n}\left(\frac{N-n}{N}\right).$$

**Теорема 1.1.4.** Нехай $\hat{Y}_{st} = N\bar{y}_{st}$ – оцінка сумарного значення популяції при стратифікованому випадковому відборі. Дисперсія оцінки має вигляд:

$$D\hat{Y}_{st} = D(N\bar{y}_{st}) = N^2 D\bar{y}_{st} = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}.$$

## 1.2. Оцінювання дисперсії та довірчі межі

Нехай з кожної страти здобувається проста випадкова вибірка, тоді незміщена оцінка дисперсії $S_h^2$ змінної $y$ в страті $h$ дорівнює

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2.$$

**Теорема 1.2.1.** При *стратифікованому випадковому відборі* незміщена оцінка дисперсії $D\bar{y}_{st}$ дорівнює

$$s^2(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}.$$

Запишемо оцінку дисперсії у зручному для обчислень вигляді:

$$s^2(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h s_h^2}{N}.$$

Довірчий інтервал для середнього значення популяції $\bar{Y}$ має вигляд

$$\left(\bar{y}_{st} - \sqrt{s^2(\bar{y}_{st})} z_{1-\alpha/2};\ \bar{y}_{st} + \sqrt{s^2(\bar{y}_{st})} z_{1-\alpha/2}\right)$$

і містить невідоме середнє значення з імовірністю $1 - \alpha$,

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

Довірчий інтервал для сумарного значення популяції $N\bar{Y}$ має вигляд

$$\left(N\bar{y}_{st} - N\sqrt{s^2(\bar{y}_{st})} z_{1-\alpha/2};\ N\bar{y}_{st} + N\sqrt{s^2(\bar{y}_{st})} z_{1-\alpha/2}\right)$$

і містить невідоме сумарне значення з імовірністю $1 - \alpha$.

### 1.3. Оптимальне розміщення

При стратифікованому відборі обсяги вибірок у стратах визначає дослідник. Їх можна вибирати так, щоб мінімізувати дисперсію $D\bar{y}_{st}$ при визначених витратах $C$ на здобуття вибірки або мінімізувати витрати за визначеної величини $D\bar{y}_{st}$.

Найпростіша функція витрат має вигляд:

$$C = c_0 + \sum_{h=1}^{L} c_h n_h.$$

Для кожної страти $h$ витрати пропорційні обсягу вибірки $n_h$, але витрати у розрахунку на одну одиницю, $c_h$, можуть змінюватись від страти до страти. Вільний член $c_0$ відповідає накладним витратам. Така функція виправдана, якщо основну частину витрат складають *витрати на вимірювання кожної одиниці*. Якщо ж істотну частину витрат складають *витрати на пересування від однієї одиниці до іншої*, то, як показали практичні й теоретичні дослідження, цим шляховим витратам відповідає вираз

$$\sum_{h=1}^{L} t_h \sqrt{n_h},$$

де $t_h$ – шляхові витрати у розрахунку на одну одиницю.

**Теорема 1.3.1**. При *стратифікованому випадковому відборі* з функцією витрат

$$C = c_0 + \sum_{h=1}^{L} c_h n_h$$

дисперсія $D\bar{y}_{st}$ оцінки середнього значення популяції мінімальна, якщо обсяги вибірок $n_h$ пропорційні $N_h S_h / \sqrt{c_h}$.

Доведення. Задача полягає у мінімізації виразу

$$D\bar{y}_{st} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

за умов, що

$$C - c_0 = \sum_{h=1}^{L} c_h n_h.$$

Застосуємо метод множників Лагранжа, виберемо $n_h$ і множник $\lambda$ такими, що мінімізують вираз:

$$D\bar{y}_{st} + \lambda\left(\sum_{h=1}^{L} c_h n_h + c_0 - C\right),$$

або, що те ж саме,

$$\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h} + \lambda\left(\sum_{h=1}^{L} c_h n_h + c_0 - C\right).$$

Продиференціюємо останній вираз за $n_h$ і прирівняємо нулеві:

$$-\frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h = 0, \qquad h = 1,2,\dots,L,$$

або,

$$n_h\sqrt{\lambda} = \frac{W_h S_h}{\sqrt{c_h}}, \qquad h = 1,2,\dots,L.$$

Підсумовуючи по всіх стратах, здобуваємо:

$$(n_1 + n_2 + \cdots + n_L)\sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}},$$

або,

$$n\sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}}.$$

Отже,

$$\frac{n_h}{n} = \frac{W_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L}\left(W_h S_h/\sqrt{c_h}\right)} = \frac{N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L}\left(N_h S_h/\sqrt{c_h}\right)}.$$

Теорема доведена.

Теорема 1.3.1 задає наступні правила відбору. Для заданої страти необхідно здобувати вибірку більшого обсягу, якщо: 1) страта велика; 2) у страті велика варіація ознаки; 3) відбір у страті виявляється дешевше.

Для того, щоб завершити розміщення, залишилось зробити ще один крок. Останнє співвідношення задає обсяги вибірок $n_h$ у частках обсягу вибірки $n$, проте ми ще не знаємо, яке значення $n$ обрати. Розв'язання останнього питання залежить від того, чи повинна вибірка забезпечити певні загальні витрати $C$ або певну дисперсію $D\bar{y}_{st}$.

Якщо *витрати C незмінні,* то необхідно підставити оптимальні значення

$$n_h = n\frac{N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L}\left(N_h S_h/\sqrt{c_h}\right)}$$

у функцію витрат

$$C = c_0 + \sum_{h=1}^{L} c_h n_h$$

і розв’язати рівняння відносно $n$. Здобуваємо:

$$n = \frac{(C - c_0)\sum_{h=1}^{L}\left(N_h S_h/\sqrt{c_h}\right)}{\sum_{h=1}^{L}\left(N_h S_h\sqrt{c_h}\right)}.$$

Якщо *дисперсія* $D\bar{y}_{st} = D$ *незмінна*, то необхідно підставити оптимальне розміщення

$$n_h = n\frac{N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L}\left(N_h S_h/\sqrt{c_h}\right)}$$

у формулу для дисперсії

$$D = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

і розв'язати рівняння відносно $n$. Здобуваємо:

$$n = \frac{\sum_{h=1}^{L}\left(W_h S_h \sqrt{c_h}\right) \sum_{h=1}^{L}\left(W_h S_h / \sqrt{c_h}\right)}{D + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}.$$

Важливий частинний випадок виникає, коли витрати у розрахунку на одиницю в усіх стратах однакові. Тоді загальні витрати набувають вигляду:

$$C = c_0 + cn,$$

і оптимальне розміщення при незмінних витратах зводиться до оптимального розміщення при незмінному обсязі вибірки. У цьому випадку теорема 1.3.1 набуває наступного вигляду.

**Теорема 1.3.2.** При *стратифікованому випадковому відборі* з незмінним обсягом вибірки $n$ дисперсія $D\bar{y}_{st}$ мінімальна, якщо

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}.$$

Таке розміщення називають *неймановим розміщенням*.

Підставимо розміщення Неймана у формулу для дисперсії $D$ і розв'яжемо рівняння відносно $n$. Здобуваємо:

$$n = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{D + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}.$$

Якщо витрати у розрахунку на одиницю в усіх стратах однакові і дисперсії $S_h^2$ в усіх стратах однакові, то оптимальне розміщення зводиться до *пропорційного розміщення*:

$$n_h = n\frac{N_h}{\sum_{h=1}^{L} N_h} = n\frac{N_h}{N}.$$

Підставимо пропорційне розміщення у формулу для дисперсії $D$ і розв'яжемо рівняння відносно $n$. Здобуваємо:

$$n = \frac{\sum_{h=1}^{L} W_h S_h^2}{D + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}.$$

## 1.4. Оцінювання часток

Якщо необхідно оцінити частку одиниць популяції, що відносяться до певного визначеного класу $C$, то ідеальна стратифікація полягає в тому, щоб включити в першу страту всі одиниці з класу $C$, а в другу — всі інші одиниці популяції. Але не маючи такої можливості, ми намагатимемося утворити страти так, щоб частка одиниць з класу $C$ змінювалась від страти до страти якомога більше.

Нехай $A_h$ – число одиниць класу $C$ у страті $h$ в популяції, $a_h$ – число одиниць класу $C$ у вибірці зі страти $h$.

Нехай

$$P_h = \frac{A_h}{N_h}, \qquad p_h = \frac{a_h}{n_h}$$

частки одиниць із класу $C$ у страті $h$ та в вибірці з цієї страти відповідно.

Природною оцінкою частки одиниць популяції при стратифікованому випадковому відборі є

$$p_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h p_h.$$

**Теорема 1.4.1.** При стратифікованому випадковому відборі дисперсія оцінки частки одиниць популяції має вигляд:

$$Dp_{st} = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2(N_h - n_h)}{N_h - 1}\frac{P_h Q_h}{n_h}, \quad Q_h = 1 - P_h.$$

Доведення. Теорема є частинним випадком теореми 1.1.3 про дисперсію оцінки середнього значення популяції. Згідно з теоремою 1.1.3:

$$D\bar{y}_{st} = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}.$$

Нехай $y_{hi}$ – змінна, яка набуває значення 1, якщо одиниця належить класу $C$, і значення 0 в супротивному випадку. Для такої змінної дисперсія в страті $h$ популяції дорівнює

$$S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2 = \frac{1}{N_h - 1}\left(\sum_{i=1}^{N_h} y_{hi}^2 - N_h \bar{Y}_h{}^2\right) =$$

$$= \frac{1}{N_h - 1}(N_h P_h - N_h P_h^2) = \frac{N_h}{N_h - 1} P_h Q_h.$$

Звідси випливає твердження теореми.

**Теорема 1.4.2.** При *стратифікованому випадковому відборі* незміщена оцінка дисперсії $Dp_{st}$ дорівнює

$$s^2(p_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{p_h q_h}{n_h - 1}, \quad q_h = 1 - p_h.$$

Довірчий інтервал для частки одиниць популяції має вигляд

$$\left(p_{st} - \sqrt{s^2(p_{st})}z_{1-\alpha/2};\ p_{st} + \sqrt{s^2(p_{st})}z_{1-\alpha/2}\right)$$

і містить невідому частку з імовірністю $1 - \alpha$, де $z_{1-\alpha/2}$ – $(1 - \alpha/2)$ –квантиль стандартного нормального розподілу.

Правила вибору обсягів вибірок $n_h$ у стратах так, щоб мінімізувати дисперсію $Dp_{st}$ при визначених витратах $C$ випливають з теорії, викладеної в підрозділі 1.1.

## 2. Дослідження телевізійної аудиторії

Надійні, повні та якісні дані забезпечують планування телевізійної сітки та реклами відповідно до вподобань та звичок телеглядачів.

ТВ-панель – це система кількісних репрезентативних панельних досліджень динамічного сегменту суспільства – ТВ аудиторії, які фіксують щосекундне дивлення ТВ за допомогою високотехнологічних пристроїв. Завдяки тому, що обрані домогосподарства за своїми соціально-демографічними і технічними характеристикам (тип прийняття телесигналу, телевізійне обладнання) відтворюють структуру населення країни, вони формують розгорнуту картину телепереглядуу всіма глядачами країни.

Робота ТВ-панелі здійснюється за допомогою спеціальної технології. У кожному домогосподарстві, яке бере участь в дослідженні, встановлюється

спеціальний прилад – піплметр. Він приєднується до кожного телевізора і з його допомогою перегляд кожного члена домогосподарства реєструється.

Клієнтами ТВ-панелі є телеканали, рекламні агенції та рекламодавці. Дані дослідження телевізійної аудиторії, здобуті в результаті роботи ТВ-панелі, – важливий інструмент оцінки ефективності програмування ТВ-каналів, аналізу аудиторії програм та рекламних повідомлень.

З 2002 по 2013 роки оператором ТВ панелі була міжнародна дослідницька компанія GfK Ukraine. До 2013 року учасниками ТВ панелі були 2,5 тис. домогосподарств чи 6,5 тис. глядачів у віці від 4 років, які мешкають в 354 населених пунктах по всій Україні. З 2014 року Індустріальний Телевізійний Комітет (ІТК) надає ринку дані дослідження телевізійної аудиторії (ТВ панель), яке по замовленню ІТК готують дослідницькі компанії Nielsen та Комунікаційний Альянс. Телевізійна панель базується на загальнонаціональній вибірці з 3740 домогосподарств, де 2540 домогосподарств розташовані в містах з населенням більше 50 тис. людей і 1200 домогосподарств в містах з населенням менше 50 тис. людей та селах [7].

**Приклад 2.1.** Нехай популяція телевізійної аудиторії складається з 310 домогосподарств (дмг), що мають принаймні один телевізор, проживають в містах А, В та сільській місцевості та мають різні соціально-демографічні характеристики. У складі 155 дмг міста А висока частка дітей шкільного віку, у складі 62 дмг міста В висока частка осіб пенсійного віку, в сільській місцевості розташовано 93 дмг. Проста випадкова вибірка дослідження телевізійної аудиторії становить 40 дмг: 20 дмг з міста А, 8 дмг з міста В та 12 дмг із сільської місцевості. Оцінити середнє та сумаре число годин телепереглядy щотижня, побудувати довірчі інтервали для середнього та сумарного числа годин телепереглядy з коефіцієнтом надійності 0,95.

Місто А: 35, 43, 36, 39, 28, 28, 29, 25, 38, 27, 26, 32, 29, 40, 35, 41, 37, 31, 45, 34.

Місто В: 27, 15, 4, 41, 49, 25, 10, 30.

Сільська місцевість: 8, 14, 12, 15, 30, 32, 21, 20, 34, 7, 11, 24.

Розв'язання. Популяція телевізійної аудиторії природно розділяється на три страти: місто А, місто В та сільська місцевість. Ваги страт дорівнюють

$$W_1 = \frac{N_1}{N} = \frac{155}{310} = 0{,}5; \quad W_2 = \frac{N_2}{N} = \frac{62}{310} = 0{,}2; \quad W_3 = \frac{N_3}{N} = \frac{93}{310} = 0{,}3.$$

Вибіркові середні в кожній страті дорівнюють

$$\bar{y}_1 = 33{,}9; \quad \bar{y}_2 = 25{,}125; \quad \bar{y}_3 = 19.$$

Незміщена оцінка середнього числа годин перегляду ТВ щотижня дорівнює

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h = \frac{155}{310} \cdot 33{,}9 + \frac{62}{310} \cdot 25{,}125 + \frac{93}{310} \cdot 19 = 27{,}7.$$

Вибіркові дисперсії в кожній страті дорівнюють

$$s_1^2 = 5{,}95^2; \quad s_2^2 = 15{,}25^2; \quad s_3^2 = 9{,}36^2.$$

Незміщена оцінка дисперсії дорівнює

$$s^2(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h s_h^2}{N} = 1{,}97.$$

Довірчим інтервалом для середнього числа годин перегляду ТВ щотижня з коефіцієнтом надійності $1 - \alpha = 0{,}95$ є інтервал

$$\left(\bar{y}_{st} - z_{1-\alpha/2}\sqrt{s^2(\bar{y}_{st})};\ \bar{y}_{st} + z_{1-\alpha/2}\sqrt{s^2(\bar{y}_{st})}\right),$$

$$\left(27{,}7 - 1{,}96\sqrt{1{,}97};\ 27{,}7 + 1{,}96\sqrt{1{,}97}\right),$$

$$(27{,}7 - 2{,}8;\ 27{,}7 + 2{,}8),$$

отже, похибка оцінки середнього числа годин складає 2,8 години.

Незміщена оцінка сумарного числа годин перегляду ТВ щотижня дорівнює

$$\hat{Y}_{st} = N\bar{y}_{st} = 310 \cdot 27{,}7 = 8587.$$

Незміщена оцінка дисперсії дорівнює

$$s^2\left(\hat{Y}_{st}\right) = N^2 s^2(\bar{y}_{st}) = 310^2 \cdot 1{,}97 = 189317.$$

Довірчим інтервалом для сумарного числа годин перегляду ТВ щотижня з коефіцієнтом надійності $1 - \alpha = 0{,}95$ є інтервал

$$\left(N\bar{y}_{st} - z_{1-\alpha/2}N\sqrt{s^2(\bar{y}_{st})};\ N\bar{y}_{st} + z_{1-\alpha/2}N\sqrt{s^2(\bar{y}_{st})}\right),$$

$$\left(8587 - 1{,}96\sqrt{189317};\ 8587 + 1{,}96\sqrt{189317}\right),$$

$$(8587 - 852{,}8;\ 8587 + 852{,}8),$$

отже, похибка оцінки сумарного числа годин складає 852,8 години.

**Приклад 2.2.** Нехай витрати на обстеження одного домогосподарства (персональне інтерв'ю) в сільській місцевості складають 16 USD, а в містах А та В – 9 USD. Дисперсії числа годин перегляду ТВ щотижня в містах А, В і сільській місцевості відомі й дорівнюють $S_1^2 = 25, S_2^2 = 225, S_3^2 = 100$. Визначити обсяг вибірки, який необхідно здобути для оцінювання середнього числа годин перегляду ТВ щотижня з похибкою, що складає дві години. Розподілити знайдений обсяг вибірки за стратами.

Розв'язання. Похибка оцінки середнього числа годин перегляду ТВ дорівнює

$$e = z_{1-\alpha/2}\sqrt{D\bar{y}_{st}},$$

звідси дисперсія дорівнює

$$D = D\bar{y}_{st} = \left(e/z_{1-\alpha/2}\right)^2 = (2/1{,}96)^2 = 1{,}04.$$

Обсяг вибірки при оптимальному розміщенні дорівнює

$$n = \frac{\sum_{h=1}^{L}\left(W_h S_h\sqrt{c_h}\right)\sum_{h=1}^{L}\left(W_h S_h/\sqrt{c_h}\right)}{D + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2},$$

зокрема, маємо

$$\sum_{h=1}^{L}\left(W_h S_h\sqrt{c_h}\right) = 0{,}5\cdot 5\cdot\sqrt{9} + 0{,}2\cdot 15\cdot\sqrt{9} + 0{,}3\cdot 10\cdot\sqrt{16} = 28{,}5,$$

$$\sum_{h=1}^{L}\left(W_h S_h/\sqrt{c_h}\right) = 0{,}5\cdot 5/\sqrt{9} + 0{,}2\cdot 15/\sqrt{9} + 0{,}3\cdot 10/\sqrt{16} = 2{,}58,$$

$$\sum_{h=1}^{L} W_h S_h^2 = 0{,}5\cdot 25 + 0{,}2\cdot 225 + 0{,}3\cdot 100 = 87{,}5,$$

отже, обсяг вибірки, який необхідно здобути для оцінювання середнього числа годин перегляду ТВ щотижня з похибкою, що складає дві години, дорівнює 56.

Розподіляємо здобутий обсяг вибірки за стратами:

$$n_1 = n\frac{N_1 S_1/\sqrt{c_1}}{\sum_{h=1}^{3}\left(N_h S_h/\sqrt{c_h}\right)} = 56\cdot\frac{155\cdot 5/\sqrt{9}}{155\cdot 5/\sqrt{9} + 62\cdot 15/\sqrt{9} + 93\cdot 10/\sqrt{16}} = 18,$$

$$n_2 = n\frac{N_2 S_2/\sqrt{c_2}}{\sum_{h=1}^{3}\left(N_h S_h/\sqrt{c_h}\right)} = 56\cdot\frac{62\cdot 15/\sqrt{9}}{155\cdot 5/\sqrt{9} + 62\cdot 15/\sqrt{9} + 93\cdot 10/\sqrt{16}} = 22,$$

$$n_3 = n\frac{N_3 S_3/\sqrt{c_3}}{\sum_{h=1}^{3}\left(N_h S_h/\sqrt{c_h}\right)} = 56\cdot\frac{93\cdot 10/\sqrt{16}}{155\cdot 5/\sqrt{9} + 62\cdot 15/\sqrt{9} + 93\cdot 10/\sqrt{16}} = 16.$$

*Зауваження.* Для знаходження оптимальних значень $n_1, n_2, \dots, n_L$ необхідно знати дисперсії $S_1^2, S_2^2, \dots, S_L^2$. Для цього на практиці необхідно обстежити всю популяцію, що часто неможливо. Але, якщо відомі результати пілотного обстеження, то можна наближено оцінити дисперсії $S_1^2, S_2^2, \dots, S_L^2$. Тоді дістанемо розміщення близьке до оптимального.

**Приклад 2.3.** Нехай витрати на обстеження одного домогосподарства (телефонне інтерв'ю) в містах А, В та сільській місцевості однакові. Дисперсії числа годин перегляду ТВ щотижня в містах А, В і сільській місцевості відомі й дорівнюють $S_1^2 = 25, S_2^2 = 225, S_3^2 = 100$. Визначити обсяг вибірки, який необхідно здобути для оцінювання середнього числа годин перегляду ТВ щотижня з похибкою, що складає дві години. Розподілити знайдений обсяг вибірки за стратами.

Розв'язання. Похибка оцінки середнього числа годин перегляду ТВ дорівнює

$$e = z_{1-\alpha/2}\sqrt{D\bar{y}_{st}},$$

звідси дисперсія дорівнює

$$D = D\bar{y}_{st} = \left(e/z_{1-\alpha/2}\right)^2 = (2/1{,}96)^2 = 1{,}04.$$

Обсяг вибірки при нейманновому розміщенні дорівнює

$$n = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{D + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2},$$

зокрема, маємо

$$\sum_{h=1}^{L} W_h S_h = 0{,}5 \cdot 5 + 0{,}2 \cdot 15 + 0{,}3 \cdot 10 = 8{,}5,$$

$$\sum_{h=1}^{L} W_h S_h^2 = 0{,}5 \cdot 25 + 0{,}2 \cdot 225 + 0{,}3 \cdot 100 = 87{,}5,$$

отже, обсяг вибірки, який необхідно здобути для оцінювання середнього числа годин перегляду ТВ щотижня з похибкою, що складає дві години, дорівнює 55.

Розподіляємо здобутий обсяг вибірки за стратами:

$$n_1 = n\frac{N_1 S_1}{\sum_{h=1}^{3} N_h S_h} = 55 \cdot \frac{155 \cdot 5}{155 \cdot 5 + 62 \cdot 15 + 93 \cdot 10} = 17,$$

$$n_2 = n\frac{N_2 S_2}{\sum_{h=1}^{3} N_h S_h} = 55 \cdot \frac{62 \cdot 15}{155 \cdot 5 + 62 \cdot 15 + 93 \cdot 10} = 19,$$

$$n_3 = n\frac{N_3 S_3}{\sum_{h=1}^{3} N_h S_h} = 55 \cdot \frac{93 \cdot 10}{155 \cdot 5 + 62 \cdot 15 + 93 \cdot 10} = 19.$$

**Приклад 2.4.** Нехай витрати на обстеження одного домогосподарства (телефонне інтерв'ю) в містах А, В та сільській місцевості однакові. Дисперсії числа годин перегляду ТВ щотижня в містах А, В і сільській місцевості відомі й однакові $S_1^2 = S_2^2 = S_3^2 = 100$. Визначити обсяг вибірки, який необхідно здобути для оцінювання середнього числа годин перегляду ТВ щотижня з похибкою, що складає дві години. Розподілити знайдений обсяг вибірки за стратами.

Розв'язання. Похибка оцінки середнього числа годин перегляду ТВ дорівнює

$$e = z_{1-\alpha/2}\sqrt{D\bar{y}_{st}},$$

звідси дисперсія дорівнює

$$D = D\bar{y}_{st} = \left(e/z_{1-\alpha/2}\right)^2 = (2/1{,}96)^2 = 1{,}04.$$

Обсяг вибірки при пропорційному розміщенні дорівнює

$$n = \frac{\sum_{h=1}^{L} W_h S_h^2}{D + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2},$$

зокрема, маємо

$$\sum_{h=1}^{L} W_h S_h^2 = 0{,}5 \cdot 100 + 0{,}2 \cdot 100 + 0{,}3 \cdot 100 = 100,$$

отже, обсяг вибірки, який необхідно здобути для оцінювання середнього числа годин перегляду ТВ щотижня з похибкою, що складає дві години, дорівнює 73.

Розподіляємо здобутий обсяг вибірки за стратами:

$$n_1 = n\frac{N_1}{\sum_{h=1}^{3} N_h} = n\frac{N_1}{N} = 73 \cdot \frac{155}{310} = 36,$$

$$n_2 = n\frac{N_2}{\sum_{h=1}^{3} N_h} = n\frac{N_2}{N} = 73 \cdot \frac{62}{310} = 15,$$

$$n_3 = n\frac{N_3}{\sum_{h=1}^{3} N_h} = n\frac{N_3}{N} = 73 \cdot \frac{93}{310} = 22.$$

**Приклад 2.5.** Нехай популяція телевізійної аудиторії складається з 310 домогосподарств (дмг), що мають принаймні один телевізор, проживають в містах А, В та сільській місцевості та мають різні соціально-демографічні характеристики. У складі 155 дмг міста А висока частка дітей шкільного віку, у складі 62 дмг міста В висока частка осіб пенсійного віку, в сільській місцевості розташовано 93 дмг. Проста випадкова вибірка дослідження телевізійної аудиторії становить 40 дмг: 20 дмг з міста А, 8 дмг з міста В та 12 дмг із сільської місцевості. Підчас інтерв'ю виявилося, що шоу Х-фактор дивляють 16 дмг з міста А, 2 дмг з міста В та 6 дмг із сільської місцевості. Оцінити частку домогосподарств, в яких дивляться шоу, побудувати довірчий інтервал з коефіцієнтом надійності 0,95.

Розв'язання. Незміщена оцінка частки домогосподарств, в яких дивляться шоу Х-фактор дорівнює

$$p_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h p_h = \frac{1}{310}\left(155 \cdot \frac{16}{20} + 62 \cdot \frac{2}{8} + 93 \cdot \frac{6}{12}\right) = 0{,}65.$$

Знайдемо оцінки дисперсій часток домогосподарств, в яких дивляться шоу Х-фактор, в кожній страті

$$s^2(p_1) = \left(1 - \frac{n_1}{N_1}\right)\frac{p_1 q_1}{n_1 - 1} = \left(1 - \frac{20}{155}\right)\frac{0{,}8 \cdot 0{,}2}{20 - 1} = 0{,}007,$$

$$s^2(p_2) = \left(1 - \frac{n_2}{N_2}\right)\frac{p_2 q_2}{n_2 - 1} = \left(1 - \frac{8}{62}\right)\frac{0{,}25 \cdot 0{,}75}{8 - 1} = 0{,}023,$$

$$s^2(p_3) = \left(1 - \frac{n_3}{N_3}\right)\frac{p_3 q_3}{n_3 - 1} = \left(1 - \frac{12}{93}\right)\frac{0{,}5 \cdot 0{,}5}{12 - 1} = 0{,}020.$$

Незміщена оцінка дисперсії дорівнює

$$s^2(p_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 s^2(p_h) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{p_h q_h}{n_h - 1} = 0{,}0045.$$

Довірчим інтервалом частки домогосподарств, в яких дивляться шоу Х-фактор, з коефіцієнтом надійності $1 - \alpha = 0{,}95$ є інтервал

$$\left(p_{st} - z_{1-\alpha/2}\sqrt{s^2(p_{st})};\ p_{st} + z_{1-\alpha/2}\sqrt{s^2(p_{st})}\right),$$

$$\left(0{,}65 - 1{,}96\sqrt{0{,}0045};\ 0{,}65 + 1{,}96\sqrt{0{,}0045}\right), \qquad (0{,}65 - 0{,}13;\ 0{,}65 + 0{,}13),$$

отже, похибка оцінки частки домогосподарств складає 0,13.

**Приклад 2.6.** Нехай витрати на обстеження одного домогосподарства (персональне інтерв'ю) в сільській місцевості складають 16 USD, а в містах A та B – 9 USD. Підчас пілотного обстеження виявилося, що частки дмг, в яких дивляться шоу Х-фактор становлять $p_1 = 0{,}8, p_2 = 0{,}25, p_3 = 0{,}5$ відповідно. Обчислити обсяг вибірки, який необхідно здобути для оцінювання частки домогосподарств, в яких дивляться шоу, з похибкою 0,1. Розподілити знайдений обсяг вибірки за стратами.

Розв'язання. Похибка оцінки частки домогосподарств становить

$$e = z_{1-\alpha/2}\sqrt{Dp_{st}},$$

звідси дисперсія дорівнює

$$D = Dp_{st} = \left(e/z_{1-\alpha/2}\right)^2 = (0{,}1/1{,}96)^2 = 0{,}0026.$$

Обсяг вибірки при оптимальному розміщенні дорівнює

$$n = \frac{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{p_h q_h}{a_h}}{D + \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{p_h q_h}{N_h}},$$

де $a_h$ – частка одиниць, розміщених в страті $h$.

Розміщення, яке мінімізує витрати $C$ при заданій точності оцінки дисперсії дорівнює

$$n_h = n\left(\frac{N_h\sqrt{p_h q_h / c_h}}{\sum_{h=1}^{L} N_h\sqrt{p_h q_h / c_h}}\right),$$

отже,

$$n_1 = n\frac{155\sqrt{0{,}8 \cdot 0{,}2/9}}{155\sqrt{0{,}8 \cdot 0{,}2/9} + 62\sqrt{0{,}25 \cdot 0{,}75/9} + 93\sqrt{0{,}5 \cdot 0{,}5/16}} = n \cdot 0{,}568,$$

$$n_2 = n\frac{62\sqrt{0{,}25 \cdot 0{,}75/9}}{155\sqrt{0{,}8 \cdot 0{,}2/9} + 62\sqrt{0{,}25 \cdot 0{,}75/9} + 93\sqrt{0{,}5 \cdot 0{,}5/16}} = n \cdot 0{,}246,$$

$$n_3 = n\frac{93\sqrt{0{,}5 \cdot 0{,}5/16}}{155\sqrt{0{,}8 \cdot 0{,}2/9} + 62\sqrt{0{,}25 \cdot 0{,}75/9} + 93\sqrt{0{,}5 \cdot 0{,}5/16}} = n \cdot 0{,}185,$$

Отже, $a_1 = 0{,}568,\ a_2 = 0{,}246,\ a_3 = 0{,}185$. Обсяг вибірки дорівнює 69.

Розподіляємо здобутий обсяг вибірки за стратами:

$$n_1 = 69 \cdot 0{,}568 = 39,$$
$$n_2 = 69 \cdot 0{,}246 = 17,$$
$$n_3 = 69 \cdot 0{,}185 = 13.$$

**Приклад 2.7.** Нехай витрати на обстеження одного домогосподарства (телефонне інтерв'ю) в містах А, В та сільській місцевості однакові. Підчас пілотного обстеження виявилося, що частки дмг, що дивляться шоу Х-фактор становлять $p_1 = 0{,}8, p_2 = 0{,}25, p_3 = 0{,}5$ відповідно. Обчислити обсяг вибірки, який необхідно здобути для оцінювання частки домогосподарств, в яких дивляться шоу Х-фактор, з похибкою 0,1. Розподілити знайдений обсяг вибірки за стратами.

Розв'язання. Похибка оцінки частки домогосподарств становить

$$e = z_{1-\alpha/2}\sqrt{Dp_{st}},$$

звідси дисперсія дорівнює

$$D = Dp_{st} = \left(e/z_{1-\alpha/2}\right)^2 = (0{,}1/1{,}96)^2 = 0{,}0026.$$

Обсяг вибірки при нейманновому розміщенні дорівнює

$$n = \frac{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{p_h q_h}{a_h}}{D + \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \frac{p_h q_h}{N_h}},$$

де $a_h$ – частка одиниць, розміщених в страті $h$.

Розміщення, яке мінімізує витрати $C$ при заданій точності оцінки дисперсії дорівнює

$$n_h = n\left(\frac{N_h\sqrt{p_h q_h}}{\sum_{h=1}^{L} N_h\sqrt{p_h q_h}}\right),$$

отже,

$$n_1 = n\frac{155\sqrt{0{,}8\cdot 0{,}2}}{155\sqrt{0{,}8\cdot 0{,}2} + 62\sqrt{0{,}25\cdot 0{,}75} + 93\sqrt{0{,}5\cdot 0{,}5}} = n\cdot 0{,}458,$$

$$n_2 = n\frac{62\sqrt{0{,}25\cdot 0{,}75}}{155\sqrt{0{,}8\cdot 0{,}2} + 62\sqrt{0{,}25\cdot 0{,}75} + 93\sqrt{0{,}5\cdot 0{,}5}} = n\cdot 0{,}198,$$

$$n_3 = n\frac{93\sqrt{0{,}5\cdot 0{,}5}}{155\sqrt{0{,}8\cdot 0{,}2} + 62\sqrt{0{,}25\cdot 0{,}75} + 93\sqrt{0{,}5\cdot 0{,}5}} = n\cdot 0{,}344,$$

Отже, $a_1 = 0{,}458,\ a_2 = 0{,}198,\ a_3 = 0{,}344$. Обсяг вибірки дорівнює 59.

Розподіляємо здобутий обсяг вибірки за стратами:

$$n_1 = 59\cdot 0{,}458 = 27,$$
$$n_2 = 59\cdot 0{,}198 = 12,$$
$$n_3 = 59\cdot 0{,}344 = 20.$$

# Список рекомендованої літератури

1. J. G. Bethlehem, *Applied survey methods : a statistical perspective*, *John Wiley & Sons*, New York, 2009.

2. W. G. Cochran, *Sampling Techniques*, *John Wiley & Sons*, New York, 1977.

3. P.S. Levy, S. Lemeshov, *Sampling of Population: Methods and Applications*, *4th Edition*, *John Wiley & Sons*, New York, 2008.

4. S.L. Lohr, *Sampling: Design and Analysis*, *Cengage Learning*, 2009.

5. C.-E. Sarndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, *Springer*, New York, 1992.

6. R.L. Scheaffer, III W. Mendenhall, R.L. Ott, K. G. Gerow, *Elementary Survey Sampling*, *7th Edition*, *Cengage Learning*, 2011.

7. *ТВ панель Україна*: [Електрон. ресурс] // URL: http://tampanel.com.ua/

8. Kelly McConville, *Analyzing Survey Data in R*: [Електрон. ресурс] // URL: https://www.datacamp.com/courses/analyzing-survey-data-in-r

# Зміст