

ISSN 1977-0375

**М е т о д о л о г і ї та р о б о ч і д о к у м е н т и**

**Довідкові рекомендації вибірки дослідження**

**Введення у структуру вибірки та методи оцінки**

**видання 2008-го року**

ЄВРОПЕЙСЬКА КОМІСІЯ

**Як отримати публікації ЄС**

Наші публікації можна купити в книгарні ЄС (http://bookshop.europa.eu), в якій ви можете розмістити замовлення у торгового представника за вашим вибором.

Видавничий офіс має всесвітню мережу торгових агентів. Можна отримати їх контактні дані, надіславши факс за номером (352) 29 29-42758.

*Europe Direct - це послуга, що дозволяє отримувати відповіді на питання щодо Європейського Союзу.*

Безкоштовний номер телефону (\*):

00 800 6 7 8 9 10 11

(\*) Деякі мобільні оператори не надають доступ до номерів, що починаються на 00 800, або ці номери можуть бути платними.

Більш детальну інформацію про Європейський Союз можна отримати в Інтернеті (http://europa.eu). Люксембург: Бюро офіційних публікацій Європейських Спільнот, 2008.

ISBN 978-92-79-07840-8

ISSN 1977-0375

Cat. No. KS-RA-08-003-EN-N

**Тема:** **Загальна та регіональна статистика**

**Добірка:** **Методології та робочі документи**

© Європейські Спільноти, 2008

Видання підготовлено професором Risto Lehtonen і старшим науковим співробітником Kari Djerf. Рукопис завершений в серпні 2007 року.

ЄВРОСТАТ

L-2920 Люксембург — Тел.: (352) 43 01-1 — адреса сайту <http://ec.europa.eu/eurostat>



Євростат - Статистичне бюро Європейських Спільнот. Його місія полягає в наданні Європейському Союзу високоякісної статистичної інформації. З цією метою він збирає і аналізує показники, отримані від національних статистичних бюро по всій Європі і забезпечує зіставні та узгоджені дані для Європейського Союзу для використання у визначенні, реалізації та аналізу політик Спільноти. Його статистичні продукти та послуги також мають велике значення для бізнес-спільноти Європи, професійних організацій, наукових кіл, бібліотекарів, недержавних організацій, ЗМІ та громадян.

Програма публікацій Євростату складається з декількох зібрань:

* **Прес-релізи** надають останню інформацію стосовно Євро-показників, а також щодо соціальних, економічних, регіональних, сільськогосподарських та екологічних тематик.
* **Статистичні збірники** - є більшими публікаціями зі статистичними даними та аналізом формату А4.
* **Кишенькові довідники** - це безкоштовні публікації, спрямовані на надання користувачам цілого набору базових показників на конкретні темі.
* **Стислі статистичні показники** забезпечуються оновленими звітами щодо основних результатів спостережень, досліджень і статистичного аналізу.
* **Стислі дані** забезпечуються останніми статистичними показниками з методологічними поясненнями.
* **Методології та робочі документи** - це технічні публікації для статистичних експертів, що працюють у тій чи іншій галузі.

Публікації Євростату можна замовити через книгарню ЄС за адресою: [http://bookshop.](http://bookshop/) europa.eu.

Усі публікації можна також завантажити безкоштовно в форматі PDF з сайту Євростату за адресою [http://ec.europa.eu/eurostat.](http://ec.europa.eu/eurostat) Крім того, на ньому можна безкоштовно знайти бази даних Євростату, а також таблиці з найбільш часто використовуваними та затребуваними короткостроковими та довгостроковими показниками.

Євростат разом з членами «Європейської статистичної системи» (ЄСС) створив мережу центрів підтримки користувачів, які існують майже у всіх державах-членах, а також у деяких країнах ЄАВТ. Їх місія полягає у наданні допомоги та консультацій інтернет-користувачам, що використовують європейські статистичні дані. Контактну інформація цієї мережі підтримки можна знайти на сайті Євростату.

**Зміст**

1. [Введення](#_TOC_250020) 7
2. [Планування дослідження та звітність 7](#_TOC_250019)
	1. [Основні концепції та визначення 7](#_TOC_250018)
	2. [Загальний план дослідження 8](#_TOC_250017)
	3. [Повідомлення якості дослідження 9](#_TOC_250016)
	4. [Питання основи вибірки 10](#_TOC_250015)
3. [Методи визначення вибірки та оцінок 11](#_TOC_250014)
	1. [Підготовчі заходи 11](#_TOC_250013)
	2. [Основні методи відбору 15](#_TOC_250012)
		1. [Проста випадкова вибірка 1](#_TOC_250011)6
		2. [Систематична вибірка 1](#_TOC_250010)8
		3. [Вибірка з ймовірністю, пропорційною до розміру 19](#_TOC_250009)
		4. [Стратифікована вибірка та методи розподілу 20](#_TOC_250008)
		5. [Кластерна вибірка 23](#_TOC_250007)
	3. [Визначення розміру вибірки 2](#_TOC_250006)6
	4. Використання допоміжної інформації на стадії оцінки 29
4. [Відсутність відповіді](#_TOC_250005) 30
	1. [Повторне зважування з метою коригування відсутності відповіді від одиниці 3](#_TOC_250004)1
	2. [Підстановки для коригування відсутності відповіді від елементу 3](#_TOC_250003)3
5. [Програмне забезпечення 3](#_TOC_250002)4
6. [Список літератури](#_TOC_250001) 36
7. [Посилання на веб-ресурси](#_TOC_250000) 38

Додаток 1 39

# Введення

Ці довідкові рекомендації стосуються основних принципів та методів вибірки дослідження. До них включаються планування дослідження, здійснення вибірки та оцінка, а також неотримання відповіді. Цей підхід не є технічним; до нього входять тільки деякі необхідні технічні матеріали. Методи ілюструватимуться практичними прикладами і при необхідності здійснюватимуться посилання на статистичне програмне забезпечення.

Оскільки неможливо розглянути всі аспекти вибірки дослідження, в деяких коротких рекомендаціях ми розглядатимемо окремі теми, які на нашу думку, мають найбільше значення для читачів. Ми націлювалися на практичне керівництво, призначене для експертів, чий практичний досвід вибірки дослідження обмежений, але які мають певні базові знання щодо основних статистичних даних. Для отримання додаткової інформації щодо розглянутих тем та додаткових питань ми наводимо посилання на обрану літературу.

Рекомендації організовані наступним чином. У Главі 2 обговорюються планування обстеження та звітність. Наводиться ряд основних понять і визначень, включаючи якість дослідження. Основні методи вибірки вводяться в Главі 3. Ми обговорюємо такі методи, як проста випадкова вибірка, систематична вибірка та кластерна вибірка. Використання допоміжної інформації відіграє ключову роль для сучасної вибірки дослідження, при цьому обговорюються такі методи як вибірка PPS, стратифікована вибірка та методи на основі моделей, включаючи оцінку у вигляді співвідношення та оцінювання по рівнянню регресії. Розглядається та ілюструється визначення розміру вибірки. Глава 4 присвячена відсутності відповіді і в ній обговорюються методи повторного зважування та підстановок. В Главі 5 наводиться стислий опис наявного програмного забезпечення та аналіз. Ми включили повний список посилань на сучасну літературу з вибірки дослідження в Главі 6. Глава 7 містить список обраних посилань на веб-матеріали, що мають відношення до цієї області.

# Планування дослідження та звітність

## Основні концепції та визначення

**Визначення дослідження**

*Дослідження* означає будь-яку форму збору даних. *Вибірка дослідження* є більш обмеженим за масштабом: збір даних ґрунтується на вибірці, підмножині загальної сукупності, тобто не загальній кількості цільових сукупностей, що називається *перепис*. Проте, у вибірці дослідження деякі субсукупності можуть досліджуватися повністю, але щодо більшості субсукупностей проводиться вибірка. У наступних главах термін *дослідження* означає вибіркові дослідження.

**Описові дослідження у порівнянні з аналітичними дослідженнями**

*Описові дослідження*, в тому числі переписи, є типовими явищами у статистичних бюро. Вони, як правило, надають інформацію щодо таких параметрів, як загальні підсумки, середні значення або пропорції на рівні загальної сукупності або на рівні деяких чітко визначених субсукупностей. В дослідженнях, в яких акцент робиться на *аналізі,* інтерес зосереджений на зв’язках та взаємозалежності між явищами. Необхідні параметри пов’язані з статистичними моделями, як-от лінійні моделі, і представлені за допомогою коефіцієнтів кореляції або регресії. Тим не менш, для обох типів важливо оцінити невідомі параметри як можна надійніше.

**Соціальні дослідження у порівнянні з бізнес-дослідженнями**

В соціальних дослідженнях в центрі уваги знаходяться особи та домогосподарства: наприклад, статистика населення, участь робочої сили, зарплата, споживання домашніх господарств, розподіл бідності та доходів, освіти, культурні заходи, охорона здоров’я та інші актуальні теми.

В бізнес-дослідженнях в центрі уваги знаходяться, підприємства, установи та/або бізнес одиниці, на кшталт місцевих одиниць одного виду діяльності, в тому числі фермерські господарства. Інтерес може варіюватися від складу виробництва та обсягів інвестиційних планів, зайнятості, використання енергії, відходів тощо.

Соціальні дослідження відрізняються від бізнес-досліджень і в інших аспектах. В офіційній статистиці бізнес-дослідження часто є обов’язковими, а соціологічні опитування, як правило, є добровільними. В соціальних дослідженнях режими збору даних є більш універсальними, може відрізнятися навіть структура вибірки.

## Загальний план дослідження

В останні роки тему методології дослідження було надруковано багато підручників. Groves та ін. (2004) надають вичерпний огляд всього процесу від планування до аналізу та інтерпретації. Крім того, існує цілий ряд спеціальної літератури щодо різних режимів збору даних, анкет для тестувань та питань, стратегій інтерв’ю тощо.

Операційні фази дослідження описуються, наприклад, Sundgren (1999). Вона включає в себе різні завдання від визначення основних цілей, стратегії збору даних, обробки даних, виробничих результатів, оцінки якості до архівування. Всі завдання є важливими, щоб гарантувати різні способи використання даних та їх якість. Читачам рекомендуються отримати додаткову інформацію з відповідної літератури на кшталт Lyberg та ін. (1997), або Biemer і Lyberg (2003).



**Рисунок 1.** Блок-схема процесу дослідження (див., наприклад, Статистичне управління Фінляндії)

## Повідомлення якості дослідження

Користувачам слід повідомляти відповідну інформацію про якість дослідження, бажано на всіх стадіях процесу дослідження. Повідомляти про якість дослідження, виділяючи різні джерела помилок, які можуть виникнути під час багатьох стадій процедур дослідження, вже стало традицією. Наприклад, Biemer і Lyberg (2003) описують такі типи помилок: Помилки конкретизації, блоку даних, неотримання відповіді, вимірювання, обробки та вибірки. Деякі з них можуть виникнути випадково, але, на жаль, з різними джерелами з’являються систематичні помилки.

**Помилки вибірки**

Стандартні помилки для оцінюваних параметрів, часто точкові оцінки є найстарішими показниками якості. Вони (та інші оцінки, отримані за допомогою оцінок, схожих на коефіцієнти варіації чи інтервали довіри) були впроваджені в період підйому методології дослідження в 1940-х роках.

**Помилки вимірювання**

Окрім помилок вибірки досить рано були впроваджені інші типи помилок. ООН вже в 1950-х роках надала перші рекомендації щодо повідомлення якості дослідження, при цьому помилки вимірювання вже враховувалися. Проте, реалізація систематичних повідомлень тривала набагато довше.

**Сукупна помилка дослідження**

Сукупна помилка обстеження щодо параметру θ вимірюється середньоквадратичною помилкою (MSE), тобто суми дисперсії і систематичної помилки в квадраті:
.

Вибіркова дисперсія отримується на основі плану вибірки, а такі інші компоненти, як розмір вибірки, мінливість параметрів відсотка та ваги вибірки, впливають на її оцінку. Помилка вибірки (тобто квадратний корінь з дисперсії вибірки) за визначенням є випадковою помилкою. Систематична помилка - це різниця між істинним значенням і очікуваною оцінкою. Коли вона не дорівнює нулю, вона вказує на систематичну помилку. На жаль, оцінка MSE вимагає повторного здійснення вибірки, а отже можуть виникнути певні ускладнення при виконанні великомасштабних обстежень. Однак для оцінки сукупної помилки було запропоновано деякі «тонкі» методи (див. Леслер та Калсбик 1992).

Характеристики якості та стандарти Європейської статистичної системи забезпечують гарну основу для повідомлення якості. До характеристик якості включаються актуальність, точність, своєчасність, пунктуальність, порівнянність, узгодженість, доступність і ясність. Актуальність описує, як статистичне дослідження задовольняє потребам та вимогами користувачів. Точність містить традиційні показники щодо якості дослідження, як-от стандартна помилка, інтервал довіри та коефіцієнт варіації тощо. Своєчасність і пунктуальність вимірюють свіжість даних і результатів. Порівнянність і узгодженість пов’язані з різними формами порівняння, - різними джерелами, що описують одне явище, порівнянністю одного обстеження з точки зору різних аспектів, як-от географічний район, порівнянності у часі тощо. Нарешті, доступність і ясність описують різні способи отримання та поширення даних і результатів, метадані та інші форми підтримки користувачів.

Крім того, побудований перелік характеристик якості для того, щоб полегшити контроль для тих досліджень, які повторюються на регулярній основі.

На веб-сайті Євростату щодо якості представлені відповідні документи стосовно повідомлення якості та деяка поточна практика і рекомендації з цього питання: [http://ec.europa.eu/eurostat/quality.](http://ec.europa.eu/eurostat/quality)

Міжнародний валютний фонд (МВФ) та Організація економічного співробітництва та розвитку (ОЕСР) створили власні стандарти, які також широко використовуються, особливо в галузі економічної статистики (див. Міжнародний валютний фонд, 2003).

## Питання основи вибірки

**Сукупність та основа вибірки**

**Цільова сукупність** - це сукупність в якій ми теоретично зацікавлені. Вважається, що вона є фіксованою (і визначеною).

**Сукупність основи вибірки** - це сукупність, яку ми можемо отримати.

**Сукупність обстеження** лежить на перетині наведених вище понять.

Ці три групи сукупності збігаються не повністю, тому що сукупність основи вибірки, як правило, містить деякі помилкові елементи, які називаються помилками репрезентативності. Нижче наведені деякі типові причини помилок репрезентативності:

* + - затримка у часі між моментом створення вибіркової сукупності та моментом її фактичного використання
		- не включення в основу вибірки нових утворень
		- нездатність включити або виключити елементи, що перейшли до іншої категорії (фізична зміна адреси, підприємства, що перейшли до іншої галузі промисловості тощо)
		- подальше включення елементів, що припинили своє існування, та інших елементів, що не включаються до вибірки

**Зайве охоплення** означає, що наша основа вибірки містить елементи, які не належать до нашої цільової сукупності. Зайве охоплення, як правило, можна виявити в ході польових робіт.

**Недостатнє охоплення** є набагато більш проблематичним, оскільки часто його неможливо виявити та оцінити надійним чином.

Може не існувати реального способу включити всі можливі різниці між цільовою сукупністю та кінцевою основою вибірки, але ті, що вже відомі, необхідно включити. Kish (1965) виступав за страту сюрпризів, яка б включала такі випадки.

Іноді для цільової сукупності не існує належної основи вибірки і тому необхідно шукати інші рішення, описані нижче.

**Кілька основ вибірки**

Кілька основ вибірки можуть мати місце, коли цільову сукупності можна отримати з кількох незалежних джерел. Використання кількох основ вибірки не є рідкістю для країн, що розвиваються, але при цьому вони також можуть використовуватися в розвинених суспільствах при дослідженні нових явищ.

**Кластерні основи вибірки**

Цілком може статися так, що для кінцевих одиниць вибірки можуть бути відсутні гарні основі вибірки сукупності, або їх створення є занадто дорогим. В цьому випадку можна застосувати альтернативу, що складається з комбінацій елементів, тобто застосувати кластерні основи вибірки.

Візьмемо, наприклад, дослідження щодо школярів: навіть при наявності основи вибірки сукупності, яка б охоплювала всіх дітей, що відвідують школу, дослідження на місцевості можна зробити значно дешевшими, якщо відбирати школи та/або класи замість випадкового відбору учнів у всій сукупності.

У великих дослідженнях населення та домашніх господарств найчастіше присутні кластери, які охоплюють певні природні групи елементів: наприклад, люди, що живуть в пронумерованих районах або адміністративних регіонах тощо.

**Інші питання**

При виявленні подвійних випадків однакових елементів, їх завжди необхідно видаляти з основи вибірки.

Іноді отримати доступ до невеликих груп населення може бути неможливим, незважаючи на те, що вони відомі (наприклад, люди, що живуть у віддалених гірських селах). Для цілей зменшення вартості та з інших причин вони можуть бути видалені з вибірки. В такому випадку постає інше складне питання: чи відбивають оцінки щодо меншої сукупності властивості субсукупностей, які були видалені з вибірки?

Вибірки відсікання є іншим прикладом цієї ж проблеми. Зазвичай вибірки відсікання застосовуються в бізнес-спостереженнях, в яких найменші одиниці не сприяють отриманню необхідних властивостей щодо необхідного параметра. Однак, у зв’язку з свідомим виключенням однієї частині з цільової сукупності існує ризик отримання систематичної помилки при оцінці.

**Додаткова інформація**

Інформація, що отримуються на основі «фонових» змінних, які використовуються або на стадії вибірки (наприклад, для створення страти чи кластерів, розрахунку розмірів тощо) або після збору даних для розрахунку ваг тощо. Іноді допоміжні дані не можна отримати з основи вибірки, але можна взяти з іншого спостереження з інших джерел, таких, як наприклад, офіційна статистика.

# Методи визначення вибірки та оцінок

## Підготовчі заходи

У вибірковому дослідженні вибірка ймовірності отримується з сукупності основи вибірки, використовуючи визначений план вибірки. Як правило, структура вибірки складається з комбінації різних методів визначення вибірки. Комплексна структура вибірки може включати кластеризацию і стратифікацію, а також кілька стадій вибірки. У простих випадках, вибірка елементів здійснюється безпосередньо на основі вибірки. У всіх випадках в процедурі вибірки використовуються кілька добре документованих методів визначення вибірки. До хороших прикладів належної літератури з методів відбору можна віднести: Kish (1965), Cochran (1977), Lohr (1999), і Lehtonen і Pahkinen (2004), які є основним джерелом для цього розділу. Корисні додаткові матеріали з вибіркового дослідження та оцінки, включаючи приклади з обчисленнями з використанням реальних даних досліджень, можна знайти в віртуальній лабораторії VLISS в вибірковому дослідженні, яка є он-лайн версією підручника Lehtonen і Pahkinen. Додаток знаходиться у вільному доступі за адресою: [http://www.math.helsinki.fi/VLISS/.](http://www.math.helsinki.fi/VLISS/) Багато відомих методів визначення вибірки можна легко реалізувати за допомогою статистичних програмних продуктів, таких як процедура SAS в SURVEYSELECT

Властивості методів відбору варіюються за ступенем статистичної ефективності та деякими практичними аспектами, такими як придатність до тієї чи іншої задачі вибірки, вимог до застосування та користі для користувача. Часто проект, час та бюджетні обмеження впливають на вибір структури вибірки для певних умов дослідження. Важливим додатковим аспектом є роль допоміжної інформації в певній процедурі вибірки. Давайте спочатку розглянемо стандартні схеми вибірки з цієї точки зору.

**Використання допоміжної інформації при виборці та оцінці**

Часто буває корисно включити допоміжну інформацію щодо сукупності до процедури вибірки. На практиці існують різні способи отримання допоміжної інформації. Наприклад, в так званих країнах регістра (наприклад, скандинавські країни), основа вибірки, використовується в підготовці офіційної статистики, часто включає додаткову інформацію щодо елементів сукупності, або ці дані отримуються з адміністративних регістрів і об’єднуються з елементами основи вибірки на мікрорівні. В інших випадках допоміжну інформацію на агрегованому рівні можна отримати із різних джерел, як-от опублікована офіційна статистика. Використання допоміжної інформації для вибірки та оцінка є додатковою особливістю при підготовці офіційної статистики. Допоміжна інформація може бути корисною при побудові ефективної структури вибірки, а також на стадії оцінки для підвищення ефективності фактичної вибірки. Щоб бути корисною, допоміжна інформація повинна бути пов’язаною з різними досліджуваними змінними.

В *простій випадковій вибірці* (SRS) вибірка отримується без використання допоміжної інформації щодо сукупності. Тому SRS забезпечує допоміжну схему при оцінці переваг в результаті використання допоміжної інформації для більш складних структур або для поліпшення ефективності оцінки для певного зразка.

Допоміжна інформація не грає ніякої ролі при стандартному застосуванні *систематичної вибірки* (SYS). Таким чином, ефективність SYS є приблизно схожою з ефективністю SRS. Така особливість також має місце, якщо елементи сукупності в основі вибірки перебувають у випадковому порядку щодо досліджуваної змінної. В методі неявна стратифікація допоміжна інформація може використовуватися у формі переліку елементів в основі вибірки. SYS може бути більш ефективною, ніж SRS, якщо існує певний взаємозв’язок між порядком елементів в основі вибірки та значеннях досліджуваної змінною.

*Вибірка з ймовірністю, пропорційною до розміру* (PPS), - це метод, в якому допоміжні дані відіграють ключову роль. Передбачається, що допоміжна змінна - це показник розміру елементу сукупності. Можна призначити різні ймовірності включення для елементів сукупності, використовуючи змінну розміру. Ефективність покращується щодо SRS, якщо між досліджуваною змінною і змінною розміру існує сильний взаємозв’язок. PPS часто використовується в бізнес-спостереженнях та в ситуаціях, коли одиниці вибірки змінюються з показником розміру.

*Типологічна вибірка* (STR) сильно залежить від використання допоміжної інформації. В STR сукупність основи вибірки спочатку ділиться на непересічні субсукупності, що мають назву *страта*, та в кожній страті здійснюється вибірка незалежно. Якщо страти є внутрішньо однорідними відносно досліджуваних змінних, тобто якщо дисперсія всередині страти досліджуваної змінної є малою, і значна частка загальної дисперсії припадає на дисперсію між стратами, то STR може бути більш ефективним, ніж SRS.

У *кластерної вибірки* (CLU), передбачається, що сукупність ділиться на природно сформовані підгрупи, що називаються кластери. Спочатку вибірка щодо груп елементів відбирається з сукупності груп елементів. На наступній стадії всі елементи відібраних груп елементів включаються до вибірки елементів (*одноетапне відбирання кластерів*), або вибірка елементів відбирається з кожного кластеру з вибірки (*двоетапне відбирання кластерів*). Якщо кластери є однорідні всередині, що зазвичай відбувається, CLU є менш ефективним, ніж SRS. *Ефект кластеризації* можна зменшити шляхом стратифікації сукупності кластерів, якість якої зазвичай підвищується.

Наведені вище методи вибірки можуть використовуватися для побудови керованого плану вибірки для певного вибіркового дослідження, за допомогою того чи іншого методу або частіше поєднання методів. У всіх методах, за винятком SRS, допоміжну інформацію у вигляді допоміжних змінних можна включити в процедуру вибірки. Відзначимо, що використання допоміжної інформації в SRS, SYS і стратифікованій вибірці вимагає наявності значень допоміжних змінних для кожного елементу сукупності. Допоміжна інформація в кластерній вибірці застосовується принаймні до групування елементів сукупності в кластери. За наявності додаткових допоміжних даних в сукупності кластерів ці дані можуть використовуватися, наприклад, для стратифікації або вибірки PPS.

Використання допоміжної інформації на стадії вибірки є типовим для *описових досліджень*, для яких кількість досліджуваних змінних є малою. Ефективність можна підвищити, якщо існує сильний зв’язок між досліджуваними змінними та допоміжними змінними.

Допоміжну інформацію можна використати для відібраної вибірки на *стадії оцінки*. Використання допоміжної інформації на стадії оцінки передбачає гнучкість: структура вибірки може залишатися простою, а на стадії оцінки використання допоміжної інформації можна адаптувати для різних досліджуваних змінних. Крім того, вимоги до допоміжних даних в стандартних методах є не такими жорсткими у порівнянні з попереднім випадком, оскільки допоміжні дані на рівні одиниць потрібні тільки для відібраних елементів, а допоміжні дані можуть бути включені на агрегованому рівні до процедури оцінки. Деякі з стандартних методів включають оцінку у вигляді співвідношення, оцінювання по рівнянню регресії та постстратіфікацію. При включенні допоміжних даних до процедури оцінки всі з цих методів використовують статистичні моделі у якості допоміжних або робочих моделей. Тому ці методи називаються *модельні.*

При оцінюванні у вигляді співвідношення та по рівнянню регресії припускається, що загальна сукупність безперервної допоміжної змінної є відомою. Допоміжна модель є лінійною моделюю регресійного типу. При оцінюванні у вигляді співвідношення модель не має вільного члена, тобто припускається, що вільний член дорівнює нулю. Ефективність можна підвищити, якщо досліджувана змінна та допоміжна змінна корелюють. Але цей метод може виявитися неефективним при наявності ненульового вільного члена у справжній моделі. При оцінюванні по рівню регресії знову застосовується модель регресійного типу, але цього разу - з вільним членом. Ефективність можна підвищити, якщо досліджувана змінна та допоміжна змінна співвідносяться.

*Постстратифікація* схожа на стратифіковану вибірку, але при цьому стратифікація виконується після визначення вибірки. Відібрана вибірка ділиться на непересічні підгрупи, що називаються *постстрата* відповідно до категоріальної або класифікованої допоміжної змінної (або кілька таких змінних), а оцінка відбувається за процедурою стратифікованої вибірки. Як і при стратифікованій виборці ефективність можна підвищити, якщо пострата буде внутрішньо однорідна відносно досліджуваної змінної. Постстратифікація часто використовується для коригування відсутності відповіді від одиниці (див. Розділ 4.1).

Таким чином, допоміжна інформація щодо сукупності можна використовувати при побудові плану вибірки, а також для покращення ефективності проведення стадії оцінки щодо певної вибірки. Як правило, ефективність оцінки можна покращити шляхом належного використання допоміжної інформації.

**Параметри, оцінки та показники якості**

Хай наш *бажаний параметр* буде фундаментальним параметром у вибірковому досліджені, *загальній сукупності* досліджуваної змінної y. У формулі для розрахунку сукупного значення *yk*  є значеннями (невідомими) досліджуваної змінної, а N - це кількість елементів в сукупності. Багато параметрів, що зазвичай використовуються при вибірковому досліджені, як-от середні величини, пропорції співвідношення коефіцієнти регресії, можуть бути виражені як функції сукупних значень. Для того, щоб отримати *оцінку* невідомої загальної сукупності T, здійснюється вибірка з сукупності та вимірюються вибіркові значення досліджуваної змінної. *Значення оцінки* щодо загальної сукупності T позначається *t*ˆ. Концепція значення оцінки означає розрахункову формулу або алгоритм, що використовуються по відношенню до вибірки з метою отримання значення оцінки. Простим прикладом є середнє вибірки, яке розраховується з використанням n вимірювань вибірки. Використовуючи середнє вибірки, оцінка загальної сукупності розраховується за наступною формулою *t*ˆ=N x ȳ. Ці розрахунки справедливі для простих структур вибірки, для більш складних структур вибірки необхідні більш складні розрахунки.

При здійсненні вибіркового дослідження віддається перевага оцінкам, які мають певні теоретичні властивості. Вони відносяться до *відсутності систематичної вибірки*, що означає, що очікування щодо значення оцінки співпадає з цільовим параметром, тобто *E*(*t*ˆ)=*T*, а систематична помилка визначається таким чином:
 *Bias*(*t*ˆ) =*E*(*t*ˆ)-*T* . *Послідовність* є дещо слабшою властивістю, і позначає поведінку значення оцінки для покращення відповідності значенню цільового параметру при збільшення розміру вибірки *n, а* також для відтворення цільового параметра, якщо розмір вибірки збігається з розміром сукупності, *N*. *Точність* значення оцінки означає її здатність змінюватися та вимірюється *дисперсією структури* *Var*(*t*ˆ) . Чим меншою є дисперсія структури, тим кращою є точність. Точна оцінка називається *ефективною*. А *точність* оцінки відноситься до об’єднаної систематичної помилки та властивостей точності щодо значення оцінки і вимірюється середньоквадратичною помилкою: *MSE*(*t*ˆ)=*Var*(*t*ˆ)+*Bias*2 (*t*ˆ) .

В практиці вибіркових досліджень використовуються значення оцінки, які є незміщеними або принаймні послідовними. Завдання статистика, що здійснює дослідження, полягає в тому, щоб при здійсненні певної задачі вибірки, отримати ефективні значення оцінок, дисперсія структури яких була б мінімальною. Це забезпечує високу надійність результатів, розрахованих за допомогою зібраних даних вибіркових досліджень.

Зазвичай у якості показників якості значень оцінки використовуються *стандартна помилка* (s.e), *варіаційний коефіцієнт* (c.v) і дизайн-ефект (deff) оцінки. Показники якості можна отримати з теоретичних властивостей, введених вище. Для значення оцінки *t*ˆ щодо загальної сукупності показники визначаються наступним способом

*Розрахункова стандартна помилка*: s.e(*t*ˆ)=, де *v*ˆ(*t*ˆ) - *розрахункова дисперсія структури* або *дисперсія вибірки* сукупного значення оцінки *t*ˆ .

*Розрахунковий варіаційний коефіцієнт або відносна стандартна помилка*:
 c.v(*t*ˆ) = s.e(*t*ˆ) / *t*ˆ , тобто розрахункова стандартна помилка, поділена на саму оцінку. Варіаційний коефіцієнт часто виражається у відсотках, 100 x c.v% . Варіаційний коефіцієнт регулярно повідомляється в офіційній статистиці. Метод C.v часто використовується як стандарт якості в контексті ЕСС (див. Розділ 3.3).

Дизайн-ефект (Deff) (Kish, 1965) вимірює *статистичну ефективність* структури вибірки по відношенню до простої випадкової вибірки (SRS) і задається наступною формулою:



де чисельник є дисперсією вибірки щодо сукупного значення вибірки відповідно до фактичної (можливо складної) структури вибірки, а знаменник являє собою дисперсію вибірки відповідно до припущення щодо простої випадкової вибірки для вибірки однакового розміру. Використовуючи дизайн-ефекту, *ефективний розмір вибірки* визначається за наступною формулою: *n*eff =*n* / deff (*t*ˆ) , тобто фактичний розмір вибірки *n*, поділений на дизайн-ефект щодо сукупного значення оцінки.

Формула для дизайн-ефекту deff вимагає наступних зауважень:

1. deff < 1 Фактична структура вибірки є *більш ефективною* ніж SRS. Відповідно, ефективний розмір вибірки є більшим, ніж фактичний розмір вибірки.
2. deff = 1 Ефективність фактичної структури вибірки є аналогічною до SRS.
3. deff > 1 Фактична структура вибірки є *менш ефективною* ніж SRS. У цьому випадку, ефективний розмір вибірки є меншим, ніж фактичний розмір вибірки.

В практиці вибіркових досліджень природна задача співпадає з випадком (а). Якщо виходити з цих зусиль, використання наявної допоміжної інформації в структурі вибірки має свої переваги. Для цієї мети часто використовуються стратифікована вибірка та вибірка PPS. Крім того, ефективність може бути підвищена на стадії оцінки шляхом включення допоміжних даних у процедуру оцінки за допомогою методів на основі моделей. В кластерній вибірці, зазвичай має місце випадку (в) через внутрішню однорідності кластерів відносно досліджуваної змінної.

## Основні методи відбору

До основних методів відбору відносяться *проста випадкова вибірка, систематична* *вибірка* та *вибірка з ймовірністю, пропорційною до розміру* (PPS). Ці методи використовуються в структурі вибірки в якості кінцевих методів вибору елементарних або *первинних одиниць вибірки* (PSU:s), а також для визначення рандомізації. Керована структура вибірки для дослідження часто передбачає стратифікацію, кластеризацію та кілька етапів вибірки. *Стратифікація* сукупності в непересічній субсукупності є популярним методом, в якому для підвищення ефективності може використовуватися допоміжна інформація. У *кластерній вибірці*, практичні аспекти вибірки та збору даних є основною мотивацією використання допоміжної інформації в структурі вибірки.

### Проста випадкова вибірка

*Проста випадкова вибірка* (SRS) часто розглядається як основна форма ймовірнісної вибірки. SRS застосовується в ситуаціях, коли відсутня попередня інформація щодо структури сукупності. Проста випадкова вибірка безпосередньо на основі сукупності основи вибірки гарантує, що кожен елемент населення має рівну ймовірність відбору. Таким чином, SRS є *структурою вибірки з рівними ймовірностями*.

В якості основного вибіркового методу проста випадкова вибірка може бути включена в якості невід’ємної частини структури вибірки. Крім того, проста випадкова вибірка встановлює основу для порівняння відносної ефективності структури вибірки за допомогою показника дизайн-ефекту, зазначеного вище.

У простій випадковій вибірці, що складається з *n* елементів, кожний елемент *k* в основі вибірки сукупності із *N* елементів має таку ж саму ймовірність, тобто, π*k=π=n* / *N* . Нагадаємо, що ймовірність включення - це ймовірність включення елемента сукупності в вибірку елементів *n.* Ймовірність включення призначається для кожного елемента сукупності перед проведенням процедури відбору проб. Ймовірності включення залежать від структури вибірки і за визначенням більше нуля для всіх елементів сукупності.

На практиці SRS може здійснюватися без заміни (SRS-WOR) або з заміною (SRS-WR). Вибірка типу WOR означає випадок, коли відібраний елемент не замінюється в сукупності. Вона також означає, що елемент сукупності може бути відібраний тільки один раз. У схемі WR відібраний елемент замінюється сукупністю. В обох випадках, залишається ймовірність включення π = *n* / *N* , а різниця полягає тільки у формулі дисперсії для бажаного статистичного показника. Як правило, SRS типу WOR є більш ефективним, SRS типу WR, тобто дисперсія SRS-WOR має тенденцію бути меншою, ніж аналогічний показник для SRS-WR. Ця властивість має місце і для інших структур вибірки і пояснює часте використання структур без заміни у практиці вибіркових досліджень.

За SRS показник цільового параметру *T* можна записати просто як



де  є середнім вибірковим. Крім того, використовуючи ймовірності включення SRS π, оцінка може бути виражена у вигляді



де *wk = N* / *n*  є *вагами вибірки*, тобто зворотною ймовірність включення. Відзначимо, що в SRS ваги вибірки є однаковими для всіх елементів вибірки. У більш складних структурах, які розглядаються, ваги вибірки можуть варіюватися від елементів (як у вибірці PPS) до груп елементів (як у стратифікованій вибірці)

Використовуючи сукупне оцінне значення, середнє значення сукупності або середнє  можна визначити за допомогою ӯ *= t*ˆ / *N* . Зверніть увагу, що ми

зробили припущення щодо відомого розміру сукупності *N*, що на практиці є реалістичним припущенням. Але якщо *N* є невідомим на стадії оцінки, показники

можна використати для розміру сукупності.

Для значення оцінки *t*ˆ щодо загальної сукупності відповідно до SRS-WOR, дисперсію вибірки *t*ˆ можна задати

*v*ˆ(*t*ˆ) =*N* 2 (1-*n* / *N* )(1/ *n*)*s*ˆ2

де  є дисперсією вибірки щодо досліджуваної змінної *y*. Кількість (1-*n* / *N* ) в формулі дисперсії вибірки називається *кінцева корекція сукупності* (fpc). Зауважимо, що якщо доля вибірки *n* / *N* є малою, як у випадку типових структур вибірки щодо осіб чи домашніх господарств, практична значимість fpc є незначною, оскільки fpc наближується до нуля. Але так ситуація не обов’язково виникатиме щодо структури вибірки в бізнес-спостереженнях, в яких частки вибірки можуть бути значно більшими.

Для SRS-WR єдина різниця в дисперсії вибірки *v*ˆ(*t*ˆ) полягає в тому, що fpc задається формулою (1-1/ *N* ) . Ця різниця також вказує на більш високу ефективність структури SRS-WOR: дизайн-ефект *t*ˆ відповідно до SRS-WR є deff (*t*ˆ) =(1-1/ *N* ) /(1-*n* / *N* ) >1 , виходячи з того, що розмір вибірки *n* є більшим одного і меншим розміру сукупності *N*. Зауважимо, що в якості допоміжної структури SRS ми використовували SRS-WOR для формули deff; це є природнім вибором, але іноді SRS-WR перебуває в цій ролі в певному статистичному програмному забезпеченні.

Підведемо підсумок: якщо доля вибірки (*n/N*) є малою, fpc для SRS-WOR буде наближатися до 1. І навпаки: якщо розмір вибірки *n* наближається до розміру сукупності *N*, оцінка дисперсії *v*ˆ(*t*ˆ) зменшуватиметься. Таким чином, для перепису дисперсія вибірки дорівнює нулю.

На практиці SRS виконується за допомогою відповідного програмного забезпечення. Наприклад, процедура SAS SURVEYSELECT може використовуватися як для SRS-WR, так і для SRS-WOR. В реальні ситуації, при використанні для вибірки SRS ми в основному маємо справу з типом структури SRS без заміни.

**Приклад.** *Вибірки Бернуллі* являє собою приклад схеми вибірки типу SRS-WOR. У цьому методі розмір вибірки не є фіксованим заздалегідь, а являється випадковою величиною, очікування щодо якої дорівнює *n*, - бажаному розміру вибірки. Це властивість призводить до зміни розміру вибірки з очікуваним значенням *N*π та дисперсією *N(*1 − π*)*π, де π позначає ймовірність включення. Випадковість розміру вибірки є відносно неважливою для великих вибірок.

Нижче описується відповідний метод. Для здійснення вибірки Бернуллі, нам потрібно виконати наступні дії:

Крок 1: Зафіксуйте значення ймовірності включення π, де 0 *<* π *<* 1, щоб очікуваний розмір вибірки дорівнював *N*π, добутку розміру сукупності та ймовірності включення. Якщо бажаний розмір вибірки дорівнює *n*, то π = *n*/*N*.

Крок 2: Додайте три змінні, наприклад, PROB, IND і UNI, в набір даних з основою вибірки. Для PROB встановлене значення рівне обраному значенню π, і IND дорівнює нулю для всіх елементів сукупності *N*. Для UNI, значення з рівномірного розподілу в діапазоні (0, 1) обирається незалежно для кожного елемента сукупності, починаючи з першого елемента. Для генерації випадкових чисел може використовуватися генератор псевдо випадкового числа.

Крок 3: Правило для включення елемента сукупності до вибірки буде наступним. *k-ий* елемент сукупності включається до вибірки, якщо UNI *<* π та відповідно, ми встановлюємо значення IND = 1 для обраного елементу (в іншому випадку, значення IND залишається нульовим).

Крок 4: Обробляйте всі елементи сукупності послідовно за допомогою Кроку 3.

Коли кроки з 1 по 4 будуть завершені, сума значень IND для основи вибірки наближатиметься (або дорівнюватиме) до бажаному розміру вибірки *n*. Елементів, що мають значення IND = 1 являють собою вибірку Бернуллі. Цю процедуру можна легко запрограмувати, наприклад в Excel, SAS або SPSS. В Додатку 1. наведений короткий приклад вибірки Бернуллі.

### Систематична вибірка

*Систематична вибірка* (SYS) є широко використовуваним вибірковим методом в ситуаціях, коли основою вибірки зазвичай є звичайна електронна (або введена вручну) база даних, як-от регістри населення, регістри фірм чи списки шкіл. SYS також є структурою ймовірнісної вибірки, оскільки включення ймовірності елементу сукупності в вибірку елементів *n*  є π = *n* / *N* .

Нижче наведені кроки відбору систематичної вибірки з *n* елементів з сукупності з *N* елементів:

* + - 1. Визначте інтервал вибірки *q* = *N*/*n*, де ціле число *q* є розрахунковим значенням.
			2. Оберіть випадкове ціле число *a* з рівною імовірністю 1/*q* між 1 та *q* (можна використовувати генератор псевдовипадкових чисел Excel, SAS, SPSS для рівномірного розподілу в діапазоні (1, q).
			3. Виберіть елементи з порядковими номерами *a*, *a* + *q, a* +2*q*, *a* +3*q*,..., *a* + (*n*−1)*q* у вибірці.

Так, для цілого числа *q* SYS дасть вибірку елементів *n*. Якщо *q* не є цілим числом, всі інтервали вибірки можуть бути встановлені на однаковому рівні довжини за винятком одного.

На практиці існує кілька способів відбору систематичної вибірки. Той, що ми навели вище являє собою приклад вибірки SYS з одним випадковим початком. В іншому випадку можна відібрати дві або *m* незалежних систематичних вибірок з використанням описаної вище процедури. В цьому випадку розмір кожної вибірки SYS дорівнює *n*/*m* елементам, а довжина інтервалу вибірки дорівнює *m* x *q*  Цей метод є підходящим, якщо оцінка дисперсії має виконуватися за допомогою так званих

*методів реплікації* (див. Wolter 2007).

Крім того, систематичну вибірку можна відібрати використовуючи елементи основи вибірки як замкнений цикл. Починаючи з довільно обраного цілого числа *A* , з [1, *N*], відбір продовжується обранням елементів *A* + *q*, *A* + 2*q*, … , до кінця основи вибірки, а потім процедура відбору продовжується з початку основи вибірки. Цикл завершується після відбору *n*  елементів. Ці випадкові методи початку дають можливість відібрати вибірку SYS із *n* елементів, при цьому при оцінці застосовується такі ж самі методи.

У статистичних програмних продуктах, таких як процедура SURVEYSELECT SAS, є вдосконалені алгоритми вибірки SYS, які використовують дробові інтервали для точного дотримання визначеного розміру вибірки *n*.

Для SYS немає відомих аналітичних значень оцінки для дисперсії структури, навіть для такого простого значення оцінки, як загальна сума. Таким чином, на практиці використовуються наближені оцінки дисперсії (див., наприклад, Wolter 2007; Lehtonen і Pahkinen 2004, Розділ 2.4).

Оцінка за допомогою систематичної вибірки залежить від знань щодо порядку сортування основи вибірки:

1. Якщо порядок сортування вибірки можна вважати випадковим по відношенню до досліджуваних змінних та всіх допоміжних змінних, оцінка за допомогою SYS відповідатиме оцінці за допомогою SRS-WOR. Таким чином, можна використовувати формули, отримані для SRS.
2. Якщо основа вибірки сортується за допоміжною змінною (або кількома такими змінними), вибірка SYS дасть вибірку, яка точно відбиватиме структуру сукупності, враховуючи змінні, що використовуються при сортуванні. Сортування основи вибірки перед вибіркою SYS називається *неявною стратифікацію*. Наприклад, в деяких випадках доречним буде відсортувати вибірку відповідно до структури регіональної сукупності. В цьому випадку систематична вибірка збереже відповідний розподіл сукупності по регіонах. Додатково розглянемо випадок, коли сукупність вже стратифікована або існує тенденція, що відбиває порядок сукупності, або існує періодична тенденція (всі ці ситуації також можуть бути реалізовані за допомогою відповідних процедур сортування). В деяких випадках періодичність може бути неефективною, особливо якщо гармонійна дисперсія збігається з вибірковим інтервалом. Оцінка з використанням неявної стратифікації відповідає оцінці з використанням стратифікованої вибірки.

Систематична вибірка, у тому числі неявна стратифікація, може бути виконана, наприклад, за допомогою процедури SURVEYSELECT SAS.

**Приклад.** Розглянемо вибірку SYS з *n* = 200 елементів з сукупності з *N* = 2000 елементів. Інтервал вибірки дорівнює *q* = *N*/*n* = 2000/200 = 10. Далі, ми відбираємо випадкове ціле число *a*  між 1 та 10, наприклад *a* . Вибірка SYS з *n* = 200 елементів складається з елементів сукупності, що мають порядкові номери 7, 17, 27,…,1997. Ймовірність включення для кожного елемента сукупності є πk = π = *n* / *N =* 200 / 2000 = 0.1 і постійна вага вибірки для вибірки елементів дорівнює *wk = w =*10 .

### Вибірка з ймовірністю, пропорційною до розміру

При *вибірці з ймовірністю, пропорційною до розміру* (PPS), ймовірність включення залежить від розміру елемента сукупності. Можна очікувати зменшення значення дисперсії, якщо показник розміру та досліджувана змінна є тісно пов’язаними. Передбачається, що значення *Zk* з допоміжним змінним показником розміру *z* є відомим для кожного елементу сукупності *k*. Типовими показниками розміру є змінні, які фізично вимірюють розмір елемента сукупності. В бізнес-обстеженнях може використовуватися, наприклад, кількість співробітників фірми у якості показника розміру, а для опитування в школах загальна кількість учнів у школі також є хорошим показником розміру. Вибірка PPS може бути дуже ефективною, особливо для оцінки сукупного значення, якщо наявний відповідний показник розміру.

У вибірці PPS ймовірність включення елемента в вибірку елементів *n* дорівнює , де  - це сума показників розміру для сукупності *N* елементів, а *pk*  називається *ймовірністю відбору з першого разу*. Для PPS ймовірності включення π*k*  варіюються залежно від елементів, отже PPS є структурою вибірки з нерівною ймовірністю.

Вибірку PPS можна відібрати з заміною або без. Розрахунком ймовірностей включення керувати легше для типу вибірки WR, оскільки сукупність залишається без змін після кожного відбору. В PPS-WOR сукупність змінюється після кожного відбору, а ймовірності включення мають бути перераховані для інших елементів.

Основні принципи оцінки відповідно за допомогою вибірки PPS наводяться тут лише коротко. Відповідно до PPS-WOR незміщене значення оцінки загальної сукупності *T* можна задати як



де *wk =* 1/ π*k*  - це вага вибірки. Показник називається *показником* *Horvitz-Thompson* (*HT*) або *показником розширення*. Показник HT є незміщеним за структурою і на практиці є дуже популярним. Оцінка дисперсії щодо розрахункового сукупного значення дорівнюватиме



де *wkl* = 1/ π*kl*. Оцінка дисперсії з використанням показника HT містить ймовірності включення другого порядку  *πkl* (тобто ймовірності включення до вибірки і елемента *k* і елемента *l*), обчислення яких є часто непрактичним, особливо для великих зразків. Тому на практиці часто використовуються наближення. Одним з варіантів є



що відповідає схемі PPS з замінами, в якій ймовірності включення другого порядку дорівнюють нулю, оскільки відбори є взаємно незалежними.

Існують різні версії схем вибірки PPS, які можна використовувати для практичних цілей. Прикладами можуть служити метод наростаючим підсумком з заміною або без заміни, *систематичний відбір PPS з нерівними ймовірностями* та *вибірка Poisson*. Наприклад, вибірка Poisson без замін нагадує вибірку Bernoulli, в якій розмір вибірки є випадковою величиною, різниця виникає тільки при розрахунку ймовірностей включення. Незважаючи на властивості розміру випадкової вибірки, вибірка Poisson іноді вважається привабливою, оскільки кількість ймовірностей включення другого порядку зводиться до π*kl =πkπl* , що спрощує розрахунок дисперсії вибірки. Книга Brewer & Hanif (1983) є гарним джерелом для різних методів PPS. Найбільш часто використовувані методи PPS реалізовані за допомогою процедури SURVEYSELECT SAS.

### Стратифікована вибірка та методи розподілу

У *стратифікованої вибірці* (STR) цільова сукупність ділиться на непересічні субсукупності, що називаються *страти*. Вони розглядаються як окремі сукупності, для яких вибірку елементів можна здійснити незалежно. В межах страти використовуються деякі з основних методів відбору, SRS, SYS або PPS, для відбору елементів вибірки. Стратифікація передбачає гнучкість, оскільки вона дає можливість застосування різних методів відбору проб для кожної страти.

Загалом, існує кілька причин популярності стратифікованої вибірки:

1. В силу адміністративних причин, багато сукупностей основи вибірки легко діляться на природні субсукупності, які можна використовувати в стратифікації. Наприклад, здійснюється визначення страти при поділу країни на регіональні адміністративні райони, які не перекриваються.
2. Стратифікація дозволяє гнучке використання в межах всієї страти допоміжної інформації для відбору проб та оцінки. Наприклад, метод PPS можна використати при відборі в межах страти, а оцінку у вигляді співвідношення або оцінювання по рівнянню регресії можна використати для відібраної вибірки, залежно від наявності додаткової допоміжної інформації в страті.
3. Стратифікація може передбачати підвищення ефективності, якщо кожна страта є однорідною з точки зору змін в досліджуваних змінних. Таким чином, дисперсія в межах страти буде невеликою, що є має позитивний вплив на ефективність.
4. При бажанні стратифікація може гарантувати репрезентативність малих субсукупностей або галузей в зразку. Це означає, що ймовірності включення для різних страт можуть бути різними. Дисперсія контролюється за допомогою так званого *методу виділення*.

У стратифікованій виборці, сукупність ділиться на *H* непересічних субсукупностей розміром із *N*1 , *N*2 ,…, *Nh* ,..., *NH*елементів, таким чином, щоб їх сума дорівнювала *N*. Для стратифікації необхідна допоміжна інформація в основі вибірки. Регіональні, демографічні та соціально-економічні змінні є типовими змінними стратифікації. Зразок відбирається незалежно від кожної страти у разі наявності розмірів вибірки страти. *n*1 , *n*2 ,…, *nh* ,..., *nH*  елементів та їх сума дорівнює *n*, загальному розміру вибірки.

Існують альтернативні стратегії визначення розмірів вибірки страти для певного дослідження. У деяких випадках спочатку фіксується загальний розмір вибірки *n*, а потім розподіляється на страти. Це характерно для тих випадків, коли самі страти не представляють інтересу (тобто підготовка статистичних даних для окремої страти не є основною метою). Якщо спостереження включає в себе підготовку даних для кожної страти (наприклад, для регіональної області або промислової групи), то важливо переконатися в достатньо великих розмірах вибірки страти. У цьому випадку спочатку визначаються розміри вибірки страти *nh*  (див. Розділ 3.3).

Найбільш поширені методи розподілу для визначення розмірів вибірки страти включають *пропорційний розподіл, рівний розподіл, оптимальний розподіл* або *розподіл* *Neyman*, а також розподіл за ступенем або *розподіл* *Bankier*. Щоб отримати уявлення про розподіл, коротко оглянемо три згадані методи (розподіл Bankier вимагає більш детальної додаткової інформації щодо розподілу сукупності в межах страти, див., наприклад, Lehtonen & Pahkinen 2004, Розділ 3.1).

*Пропорційний розподіл* є найбільш простою схемою розподілу і широко використовується на практиці. Він передбачає знання розмірів страт, оскільки доля вибірки *nh* / *Nh*  є сталою величиною для кожної страти. Кількість елементів вибірки *nh*  в страті *h* здається значенням *nh = n × Wh* , де *Wh = Nh* / *N*  є вагами страти, а *n* є визначеним загальним розміром вибірки. Пропорційний розподіл гарантує рівну частку зразка для всіх страт і передбачає структуру вибірки з рівною ймовірністю, для якої ймовірність включення π*hk = π =n* / *N*  елемента сукупності *k* в страті *h* є сталою величиною. Таким чином, вага вибірки також є сталою величиною *whk = w = N* / *n* , а структура називається *самостійно* *зваженою.*

*Рівний розподіл* забезпечує однаковий розмір вибірки *nh =n* / *H*  для кожної страти, де *H* - це кількість страт. Якщо розміри страт *Nh*  змінюються, ймовірності включення також змінюватимуться і задаються рівнянням π*hk = nh* / *Nh =n* /(*H* x *Nh* ) для елемента *k* в страті *h*. Таким чином, ваги вибірки будуть такими: *whk =H* x *Nh* / *n* . Якщо всі розміри вибірки *Nh*  є рівними, то отримаємо структуру ймовірності π*hk π = n* / *N*.

*Оптимальний розподіл* або *розподіл Neyman* є корисним, якщо стандартні відхилення сукупності *Sh*  для окремих страт щодо досліджуваної змінної *y* є відомими або у разі наявності надійних цифр. На практиці близькі наближення до істинних стандартних відхилень можуть здійснюватися на основі досвіду, накопиченого в минулих дослідженнях. Таким чином, розподіл Neyman часто використовується для безперервних бізнес-досліджень. Спочатку обчислюються розміри вибірок. Кількість одиниць вибірки *nh*  в страті *h* за нормального розподілу обчислюється як



В цьому випадку загальний розмір вибірки *n* дорівнюватиме розмірам вибірки страти. При оптимальному розподілі, велика страта або страта зі значною внутрішньою дисперсією має більше одиниць, ніж менша за розміром або більш внутрішньо однорідна страта.

Ці три схеми розподілу проілюстровані у прикладі нижче. Далі наводиться приклад розподілу за допомогою вибірки STR з додатковими обчислювальними прикладами в додатку VLISS, який є онлайн-розширенням публікації Lehtonen & Pahkinen (2004).

У стратифікованій вибірці оцінка *t*ˆ щодо загальної сукупності *Ty*  є сумою сукупних значень оцінки страти, яка задається , де  - це показник Horvitz-Thompson для сукупного значення страти *Th* . Оскільки зразки беруться незалежно від страти, дисперсія вибірки *t*ˆ є сумою дисперсій всередині страти *v*ˆ(*t*ˆ*h*), тобто . Оскільки при виборці STR дисперсія вибірки залежить тільки від дисперсії всередині страти, було б добре спробувати побудувати внутрішньо однорідні страти по відношенню до змінної дослідження *y*.

Наприклад, якщо припустити тип SRS-WOR для кожної страти, сукупне значення оцінки задається , де *Nh* / *nh*  - це вага вибірки, специфічна для страти. При застосуванні пропорційного розподілу процедуру можна спростити наступним чином

, оскільки ваги *Nh* / *nh* дорівнюють константі *N* / *n* . Таке рівняння відображає здатність пропорційного розподілу здійснювати самостійне зважування.

Оцінка за допомогою стратифікованої вибірки обговорюється більш детально в стандартних підручниках на тему вибірки: непоганими джерелами є Kish (1965) і Lohr (1999). Стратифіковану вибірку можна здійснити, наприклад, за допомогою процедури SURVEYSELECT SAS, яка передбачає кілька дискретних змінних у вигляді змінних стратифікації.

**Приклад.** В якості простого прикладу розглянемо вибірку STR за пропорційною, рівною схемами та схемою розподілу Neyman. Вибірка, стратифікована відповідно до SRS-WOR складається з *n* = 200 елементів, відібраних з сукупності з *N* = 2000 елементів (Таблиця 1).

В сукупності *H* = 5 страт. При пропорційного розподілу, вибірка 10% відбирається з кожної старти, включаючи постійні ваги вибірки *whk = w =* 2000 / 200 *=* 10 для кожного елементу вибірки. При рівному розподілі зразок з *nh =* 200 / 5 *=* 40 елементів відбирається з кожної страти, включаючи зміні ваги вибірки *whk = Nh* / *nh*  для кожної страти *h*. Для розподілу Neyman, ми припускаємо, що в нас є достовірне знання про *Sh* , стандартне відхилення сукупності (Std. Dev.), яке дорівнює *y*, і що показник є рівним для всіх страт за винятком Страти 3, стандартне відхилення якої є більшим, що вказує на більшу дисперсію досліджуваною змінної. Розміри вибірки на рівні страти розраховуються як *nh = Nh Sh* / 64000 , *h* = 1,…,5. Ця схема розподілу забезпечує більший розмір вибірки для Страти 3 і, відповідно, меншу вагу вибірки, в порівнянні з іншими стратами. Для цих страт ваги є приблизно рівними, що відображає пропорційний розподіл.

**Таблиця 1.** Схеми пропорційного, рівного та розподілу Neyman для вибірки STR, що складається з *n* = 200 елементів із сукупності з *N* = 2000 елементів.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Страта*h* | Розмір страти *Nh* | Пропорційний розподіл | Рівний розподіл | Розподіл Неймана |
| *Nh* / *N* | Розмір вибірки *nh* | Вага вибірки *whk* | Розмір вибірки *nh* | Вага вибірки *whk* | Std. Dev.*Sh* |  *Nh Sh* | Розмір вибірки *nh* | Вага вибірки *whk* |
| 1 | 500 | 0.25 | 50 | 10 | 40 | 12.5 | 20 | 10000 | 31 | 16.1 |
| 2 | 100 | 0.05 | 10 | 10 | 40 | 2.5 | 20 | 2000 | 6 | 16.7 |
| 3 | 800 | 0.40 | 80 | 10 | 40 | 20.0 | 50 | 40000 | 125 | 6.4 |
| 4 | 200 | 0.10 | 20 | 10 | 40 | 5.0 | 20 | 4000 | 13 | 15.4 |
| 5 | 400 | 0.20 | 40 | 10 | 40 | 10.0 | 20 | 8000 | 25 | 16.0 |
| Всі | 2000 | 1.00 | 200 |  | 200 |  |  | 64000 | 200 |  |

### Кластерна вибірка

Для проведення *кластерної вибірки* з сукупності кластерів спочатку відбирається вибірка кластерів (природні групи елементів сукупності, такі як кластери співробітників в установах, кластери учнів у школах і кластери людей в домашніх господарствах) використовуючи основні методи відбору проб (SRS, SYS або PPS). Крім того, сукупність кластерів можна стратифікувати перед відбором вибірки. У *кластерній вибірці в один етап* всі елементи відібраних кластерів включаються у вибірку елементів. У *кластерній вибірці в два етапи* вибірка на рівні елементу відбирається із відібраних кластерів знову з використанням обраних основних методів відбору.

Важливою перевагою кластерної вибірки є те, що основа вибірки на рівні елементів не потрібна для всієї сукупності. Єдиними вимогами є: для основ вибірки на рівні кластерів та при кластерній вибірці в два етапи - основи для відбору елементів з відібраних кластерів. Основи вибірки на рівні кластерів часто є доступними, наприклад, для установ, шкіл, сіл, ферм, районів або одиниць, схожих на райони в місті тощо. Тому допоміжна інформація в кластерній вибірці стосується не тільки групування елементів сукупності в кластери, але і властивостей необхідних кластерів при використанні стратифікації. Стратифікація є типово багатоступеневими структурами вибірки, що застосовуються, наприклад, в бізнес-обстеженнях. Наприклад, сукупність основи вибірки щодо комерційних фірм може бути стратифікована за типом галузі або розміром групи перед здійсненням вибірки для окремих фірм.

В планах кластерної вибірки в два етапи вибірка PPS іноді використовується для одиниць першого ступеня, тобто для кластерів (наприклад, регіональні підрозділи із сукупності регіонів, підприємства з комерційного регістра тощо). Структуру вибірки з рівною ймовірністю або самостійним зважуванням можна отримати, якщо елементи відбираються з відібраних кластерів з рівним розміром вибірки.

Кластерна вибірка часто обирається у зв’язку з ефективністю витрат, тобто через низьку вартість збору даних для елементу вибірки. Це особливо вірно для сукупностей, які мають велике регіональне поширення. Використовуючи кластерну вибірку, можна значно зменшити витрати на відрядження інтерв’юерів, оскільки обсяг роботи інтерв’юера можна планувати регіонально. Тому *економічна ефективність* кластерної вибірки може бути високою. Але існують певні недоліки кластерної вибірки, які стосуються статистичної ефективності. Якщо кожен кластер досить добре відображає структуру сукупності, можна досягнути ефективності відбору проб таким чином, що стандартні помилки оцінок не перевищуватимуть помилки для простої випадкової вибірки. Однак, на практиці кластери мають тенденцію бути внутрішньо однорідними, і така *однорідність всередині кластера* збільшує стандартні помилки а, отже, зменшує *статистичну ефективність*.

Кластерна вибірка обговорюється на практичній і більш технічному рівні в стандартних підручниках на тему вибірки. Непоганим прикладом є Kish (1965). Підручник Lehtonen & Pahkinen (2004, Глави 5, 7-9) пропонують кілька реальних прикладів щодо цього явища, при цьому додаткові ілюстрації можна знайти у веб-розширенні віртуальній лабораторії VLISS для вибіркового дослідження.

**Приклад.** Дослідження PISA 2000 року. Ефективність кластерної вибірки вимірюється за допомогою оцінок *дизайн-ефекту*. Статистичний показник дизайн-ефекту був введений в Розділі 3.2. Для вибіркового середнього ӯ, deff задається
deff (ӯ) = *v*ˆ(ӯ) / *v*ˆ*SRS* (ӯ) , де *v*ˆ(ӯ) є оцінкою дисперсії, що розраховується відповідно до фактичної структури кластерної вибірки, а *v*ˆ*SRS* (ӯ) є аналогом з простої випадкової вибірки. Для зразків кластерів оцінки дизайн-ефекту, як правило, більше одиниці, що вказує на меншу ефективність відносно простої випадкової вибірки. Відповідно, *ефективний розмір вибірки* зменшується: *n*eff =*n* / deff стає меншим ніж оригінальний розмір вибірки *n*, якщо deff більше одиниці. Ефективний розмір вибірки дає розмір вибірки SRS, яка забезпечує такий же рівень точності, що і фактична кластерна вибірка з розміром *n* елементів. Проілюструємо ці властивості на прикладі, який взятий з публікації Lehtonen & Pahkinen (2004, Розділ 9.4).

Дані взяті з Програми оцінки успішності міжнародних студентів (PISA) ОЕСР. Перше спостереження PISA проводилося в 2000 році в 28 країнах-членах ОЕСР і 4 країнах, що не входять в ОЕСР. Наразі ми обговоримо грамотність з читання. Ми обрали з бази даних PISA наступні країни: Бразилія, Фінляндія, Німеччина, Угорщина, Республіка Корея, Великобританія і США. Набір даних для спостережень з цих 7 країн складається в цілому з 1388 шкіл та 32 101 учнів.

В більшості країнах-членах PISA використовувалася стратифікована кластерна вибірка. Перший етап складався з вибірки окремих шкіл з систематичною вибіркою PPS. Число студентів у школі використовувалося як показник розміру при вибірці PPS. У більшості випадків перед вибіркою здійснювалася стратифікація населення. На другому етапі обиралися зразки студентів в межах відібраних шкіл з однаковою ймовірністю.

Досліджувана змінна *y* є комбінованою оцінкою грамотності студента, яка розраховується таким чином, щоб середнє значення щодо країн-учасниць становило 500, а стандартне відхилення дорівнювало 100. У Таблиці 2 наведені відібрані описові статистичні показники. На дизайн-ефект припадає зважування, стратифікація і кластеризація. Показники deff вказують на сильний ефект кластеризації для більшості країн.

Ефективні розміри вибірки для студентів розраховуються шляхом ділення числа студентів з вибірки на оцінку дизайн-ефекту. Ефективний розмір вибірки є еквівалентом розміру вибірки, необхідного для досягнення такого рівня точності в оцінці, який був би при використанні простої випадкової вибірки з сукупності студентів без будь-якої кластеризації. Якщо спостереження залежать одне від одного, як в цьому випадку, ефективний розмір вибірки зменшується: чим більшим є дизайн-ефект, тим меншим є ефективний розмір вибірки. Незважаючи на те, що номінальні розміри вибірки студентів є великими (кілька тисяч) у всіх країнах, деякі з ефективних розмірів вибірки є досить малими (всього кілька сотень). Показники дизайн-ефекту також показують, що стандартні помилки, розраховані при припущенні (помилкового) простої випадкової вибірки, буде набагато меншим, ніж показники стандартного відхилення на основі структури (правильні) для більшості країн, що має тенденцію призводити до ненадійних статистичних висновків.

**Таблиця 2.** Описова статистика для комбінованого показника грамотності в спостереженні PISA 2000 року за країнами (в алфавітному порядку).

|  |  |  |
| --- | --- | --- |
|  | Ефективний розмір вибірки студентів | Кількість спостережень в наборі даних |
| Країна | Середнє значення | Стандартна помилка | Дизайн-ефект | Студенти | Школи |
| Бразилія | 402.9 | 3.82 | 8.33 | 476 | 3961 | 290 |
| Фінляндія | 550.7 | 2.15 | 2.79 | 1600 | 4465 | 147 |
| Німеччина | 497.4 | 5.68 | 13.47 | 305 | 4108 | 183 |
| Угорщина | 485.7 | 6.02 | 20.00 | 231 | 4613 | 184 |
| Республіка Корея | 526.6 | 3.66 | 12.99 | 351 | 4564 | 144 |
| Великобританія | 531.4 | 4.08 | 14.08 | 564 | 7935 | 328 |
| Сполучені Штати | 517.0 | 5.16 | 6.93 | 354 | 2455 | 112 |
| Всі | 500.0 |  |  | 3881 | 32101 | 1388 |
| Джерело даних: База даних PISA ОЕСР, 2001. |

## Визначення розміру вибірки

Статистики, що здійснюють дослідження, часто стикаються з питанням визначення відповідного розміру вибірки у вибірковому дослідженні. Простої відповіді не існує, тому статистики мають задати питання щодо потреб, наприклад:

1. Які досліджувані змінні та параметри для оцінки є найбільш важливими?
2. Чи наявні будь-які припущення щодо розподілу (статистичного) величин дослідження?
3. Який рівень точності необхідний для оцінок параметрів?
4. Які галузі для забезпечення оцінок є найбільш важливими, і наскільки точно мають здійснюватися оцінки?
5. Чи існують будь-які конкретні питання, які мають бути враховані, наприклад, спеціальні сукупності, які мають бути охоплені, необхідний аналіз, який має бути проведений, використовувана методологія тощо?
6. Яким буде очікуваний відсоток неотримання відповіді?
7. Які фінансові та часові обмеження?

Всі ці питання (а також багато інших) повинні бути розглянуті до планування вибірки, включаючи розмір вибірки. Перша помилка практиків - це те, що вони не враховують, що розмір населення має значення. Великий розмір населення може вплинути на розмір вибірки, за винятком випадку, коли розмір вибірки не може перевищувати розмір населення. 1

**Приклад.** Дослідження робочої сили спільноти. Регламент Ради 577/98 встановлює основні принципи, яких необхідно дотримуватися при розрахунку розміру вибірки. Для простоти візьмемо тільки перший абзац Статті 3:

*Стаття 3*

*Репрезентативність вибірки*

*1.* *Для групи безробітних осіб, яка становить 5% від населення працездатного віку, відносна стандартна помилка для оцінки середньорічних показників (або для весняних оцінок у разі, якщо щорічне опитування відбувається навесні) на рівні NUTS II не повинна перевищувати 8% відповідної субсукупності.*

*Регіони з менш ніж 300 000 жителів, звільняються від цієї вимоги.*

Таким чином необхідно враховувати різні аспекти перед фактичним розрахунком:

1. Населення працездатного віку? - Часто включаються особи віком від 15 років і вище. Іноді від 15 до 74 років.
2. Частка безробітних цього населення (незважаючи на те, що молоді люди віком, скажімо менше 20 років або звільнені особи не відносяться до кількості робочої сили).
3. Дисперсія вибірки оцінюється відповідно до застосовної структури вибірки.
4. Галузі NUTS II.

**Розподіл вибірки**

Згідно Регламенту розрахунок розміру вибірки має починатися з регіонів NUTS II, тобто на рівні географічних областей. Як правило, регіони NUTS II дуже відрізняються за розміром в різних країнах і навіть всередині країн. Тому непоганим було б застосувати стратифіковану вибірку до тих даних, які складаються з регіонів NUTS II.

Наприклад, у Фінляндії є п’ять регіонів NUTS II. Чисельність населення, що складається з людей віком від 15 до 74 років в 2006 році становило:

|  |  |  |
| --- | --- | --- |
| Кількість регіону | Основний регіон | Населення (15-74) |
| 1 | Південна Фінляндія | 1,987,000 |
| 2 | Західна Фінляндія | 999,000 |
| 3 | Східна Фінляндія | 496,000 |
| 4 | Північна Фінляндія | 470,000 |
| 5 | Автономна Територія Аландських островів | 20,000 |

Оскільки найменший регіон NUTS II (№ 5) має населення менше ніж 300 тисяч, не варто починати розрахунок з нього, а краще почати з другого найменшого регіону (№ 4).

Середньорічна чисельність безробітних склала 204 000. Оскільки загальна чисельність населення складає близько 4 млн. чоловік, цей показник становить близько 5 відсотків населення - випадково саме ця частка згадується у постанові. Крім того припустимо, що частка безробітних приблизно рівна для всіх регіонів.

**Стратифікована проста випадкова вибірка**

Почнемо розрахунок розміру вибірки з простої випадкової вибірки елементів, тобто фізичних осіб. В регламенті йдеться про те, що коефіцієнт варіації не може перевищувати 8 відсотків. Тобто для кожного регіону NUTS II (за винятком останнього), необхідно дотримуватися наступної умови:



Оцінка нашого параметру, частки безробітних від працездатного населення, становить *p*ˆ = 0,05, і ми припускаємо, що вона є приблизно рівною для всіх регіонів NUTS II.

Оскільки показник безробіття є дихотомічною змінною зі значеннями [0,1] ми можемо застосувати *біноміальний розподіл* для наближеного значення дисперсії вибірки:

*v*ˆ( *p*ˆ )= *p*ˆ (1 - *p*ˆ ) / *n*

1 Тим не менш, в деяких багатонаціональних дослідженнях розміри національної вибірки можуть бути скориговані для врахування відмінностей в чисельності населення. Саме така ситуація має місце, наприклад, в EU-SILC

Його треба підставити в наведену вище формулу:



Далі, ми підносимо два компоненти в другий ступінь і отримуємо наступну формулу:

*p*ˆ (1 - *p*ˆ ) / *n =* 0.082 *p*ˆ 2

Отже,



Таким чином, розмір вибірки дорівнює приблизно 3000 одиницям. Крім того, слід враховувати й інші питання. А саме,

1. Який очікуваний відсоток неотримання відповіді?
2. Чи можемо ми застосувати структуру вибірки в даному випадку, або нам слід звернутися до якоїсь іншої структури?

Наведений вище розмір вибірки розрахований, виходячи з припущення, що відсоток отримання відповідей склав 100%. У реальному житті цей показник має бути збільшений за рахунок очікуваного відсотка відсутності відповіді та потенційних проблем недостатнього охоплення в основі вибірки. Припустимо, наприклад, що мова йде про 15% у порівнянних соціальних дослідженнях. Завищений розмір вибірки дорівнюватиме приблизно 3500 одиницям.

Наступне питання полягає в тому, чи використовувати рівний або пропорційний розподіл (див. розділ 3.2.4). *Рівний розподіл* дасть приблизно таку ж точність для всіх страт. Тому розмір вибірки для всієї країни становитиме більше ніж 14 000 чоловік (4 × 3500 = 14000 + спеціально відібраний розмір вибірки для 5-го регіону NUTS II ).

Для *пропорційного розподілу* розмір вибірки має бути закріплений на рівні регіону №4 NUTS II у відповідності до розрахунків, наведених вище. Відсоток вибірки для цього регіону становить 3500/470000, тобто близько 0,75 відсотків населення. Використовуючи той же самий відсоток вибірки, розміри вибірки для інших регіонів NUTS II становитимуть 14800, 7450, 3700 та 150. Таким чином, загальний розмір вибірки збільшиться більш ніж удвічі: 29600 - залежно від розміру вибірки останнього регіону.

Інші основні види розподілу (наприклад, розподіл Neyman або розподіл за ступенем, див. 3.2.4) не так легко застосувати в цьому прикладі, оскільки їх мета полягає в тому, щоб розподілити вибірку оптимально для всього населення, що може не відповідати вимозі постанови щодо точності для областей NUTS II.

**Структура кластера в одну стадію**

У багатьох країнах основи вибірки є застарілими або з них не можна відібрати окремих осіб чи домогосподарства безпосередньо для соціальних досліджень. В таких випадках застосовуються певні форми структур кластерної вибірки. *Кластерна вибірка в один етап* вимагає, щоб відібрані одиниці містили кінцеві одиниці вибірки, окремих людей. Одиниці вибірки першої стадії (PSU) можуть бути, наприклад, адресами, будинками або житловими одиницями. При застосуванні *структури вибірки в дві стадії*, одиниці вибірки першої стадії часто є певними адміністративними одиницями: муніципалітетами, селами або ділянками для перепису населення. Одиниці вибірки другої стадії знову можуть включати адреси, будинки або житлових одиниць, а кінцеві одиниці вибірки складаються з окремих осіб (див. розділ 3.2.5).

У кожному разі кількість первинних і вторинних одиниць вибірки має розраховуватися по-різному. Однак, основний розрахунок для структури вибірки елементів вище може слугувати в якості вхідних даних для цієї цілі також.

Для простоти ми використовуємо стратифіковану структуру кластера в одну стадію, тобто замість осіб, ми збираємо інформацію від усіх членів домогосподарств, які обираються за допомогою основи вибірки для адрес або для житлових одиниць. Далі ми припускаємо, що ми застосовуємо схему *рівного розподілу* для страти NUTS II. Для структури кластеру нам необхідний середній розмір кластерів і деякі знання щодо оцінки дизайн-ефекту. На підставі іншого спостереження ми можемо дійти висновку, що *дизайн-ефект* (див. стор. 12) кількості безробітних з структури кластера в одну стадію дорівнює приблизно 1,5. Якщо припустити, що середній розмір кластера (тобто середня кількість задовільних членів, що відповідають необхідним умовам) становить 3, а рівень відсутності, розрахований для окремих осіб, складає 15% (тобто частота відповіді τ = 0,85), можна отримати розмір вибірки наступним чином:



Використовуючи схему рівного розподілу, загальний розмір вибірки перевищуватиме 7000 домашніх господарств (4 × 1750 + спеціально обраний розмір вибірки для останнього регіону NUTS II). Таким чином, цю вимогу можна виконати за допомогою відносно великого розміру вибірки: 7000 домашніх господарств містять 21000 осіб. Вартість польових робіт звісно можна зменшити у порівнянні з стратифікованою структурою вибірки SRS.

**Бізнес дослідження**

Визначення розміру вибірки для бізнес-досліджень може виявитися набагато складнішим, ніж у попередньому прикладі. Розподіл основних досліджуваних змінних часто є надто асиметричним, що може призвести до необхідності збалансувати різні вимоги. Звичайно основний розрахунок здійснюється за тим же алгоритмом, що описаний вище, але зазвичай структура вибірки представляє собою сильно стратифіковані структури вибірки елементів. стратифікація часто проводиться за галуззю та (класифікація КДЕС) та розміром.

Одиниці, що мають найбільший вплив, часто розглядаються як визначені одиниці, і кожна з них технічно утворює власну страту. Таким чином, дисперсія вибірки відсутня (за винятком випадку, коли певні одиниці не дають відповіді). Для менших одиниць вибірка здійснюється шляхом стратифікованої систематичної випадкової вибірки, простої випадкової вибірки або вибірки PPS. При розрахунку розміру вибірки необхідно враховувати всю необхідну інформацію, а у випадку детальної стратифікації, розрахунок повинен проводитися окремо для кожної страти.

* 1. **Використання допоміжної інформації на стадії оцінки**

У сучасній практиці вибіркового дослідження допоміжна інформація часто використовується для покращення ефективності оцінки певного зразка шляхом використання *методів оцінки за сприяння моделей*. Таким чином, на додаток до структури вибірки з’являється структура оцінки. Поняття *стратегії оцінки* іноді використовується з посиланням на комбінацію структури вибірки та структури оцінки. Таблиця 3 показує приклади стратегій, включаючи стратегії на основі структури, в яких для деяких стратегій використовується допоміжна інформація, та стратегії за сприяння моделей, в яких допоміжні дані закладаються в стадії оцінки, а для деяких стратегій, також в структуру вибірки.

**Таблиця 3.** Приклади стратегій оцінки

|  |  |  |  |
| --- | --- | --- | --- |
|  | Допоміжна інформація | Допоміжна модель | Приклади стратегій (структура вибірки \* структура оцінки) |
| **Стратегії на основі структури** |
| SRS-WOR | Не використовується | Конкретні моделі відсутні | Стратегія SRS-WOR |
| SYS | Не використовується | Конкретні моделі відсутні | Стратегія SYS |
| PPS | Змінна розміру | Конкретні моделі відсутні | Стратегія PPS |
| STR | Стратифікаціязмінні | Конкретні моделі відсутні | STR \* SRS-WOR |
| **Стратегії на основі моделей** |
| Оцінка у вигляді співвідношення | Безперервний | Регресія (вільний член відсутній) | SRS-WOR\* Оцінка у вигляді співвідношення |
| Оцінювання по рівнянню регресії | Безперервний | Регресія (з вільним членом) | STR \* Оцінювання по рівнянню регресії |
| Пост-стратифікація | Дискретна | лінійна ANOVA | SYS \* Пост-стратифікація |

При оцінці на основі моделей, допоміжні дані включаються в процедуру оцінки загальної сукупності за допомогою статистичних моделей. Для цієї мети найчастіше використовуються лінійні моделі, особливо, коли залежна змінна безперервного типу. *Оцінка у вигляді співвідношення* та *оцінювання по рівнянню* регресії використовують модель лінійної регресії, в якій незалежні змінні вважаються безперервними. В оцінці у вигляді співвідношення вільний член виключається з регресійної моделі, при цьому вільний член включається до регресійної моделі, яка лежить в основі оцінювання по рівнянню регресії. На практиці, в процедуру оцінювання по рівнянню регресії можуть бути включені кілька безперервних допоміжних змінних. Для обох методів, допоміжні дані складаються з загальних сукупностей щодо однієї або кількох безперервних змінних, які можуть надходити з таких джерел як офіційна статистика.

В *пост-стратіфікаціі* у якості допоміжної моделі використовується лінійний аналіз змінної або модель ANOVA, а незалежні змінні мають дискретний тип. Допоміжні дані складаються з осередку сукупності та маргінальних частот з однієї або кількох категоріальних змінних. Перевага всіх цих методів полягає в тому, що не передбачається доступ до допоміжних даних на рівні одиниць щодо сукупності.

Параметри моделі оцінюються на основі вибірки, що досліджується, шляхом використання методу зважених найменших квадратів (WLS). У якості допоміжних моделей також можуть використовуватися нелінійні моделі. Такі випадки трапляються особливо тоді, коли досліджувані змінні є двійковими або політомічними (див., наприклад, Lehtonen, Särndal &Veijanen2005).

Оцінка у вигляді співвідношення та оцінювання по рівнянню регресії - це особливі випадки значень *оцінки узагальненої регресії* (GREG). В книзі Särndal, Swensson & Wretman (1992) наводиться основний вихідний текст для вибіркового дослідження за сприяння моделей. Методи за сприяння моделей обговорюються на більш практичному рівні у виданні Lehtonen & Pahkinen (2004). Методи проілюстровані (з обчислювальними прикладами) у додатку VLISS, яке є он-лайн версією книги.

# Відсутність відповіді

Під час збору даних відбувається втрата інформації, основна з причин якої є відсутність відповіді від одиниці. Таким чином, число відповідей є меншим, ніж вибірка, що була обрана спочатку. Якщо розмір вибірки не завищений очікуваною відсутністю відповіді, аналізи і висновки міститимуть більший ризик щодо невірних висновків, ніж це передбачалося на початку.

Відсутність відповіді від одиниці є найбільш поширеною причиною відсутніх даних. Класифікація нижче показує, як вони впливають на дані для аналізу. У підручнику у співавторстві з Groves (2002) зазначені такі ефекти, а також спроби їх врахування при аналізі.

|  |  |
| --- | --- |
| Види недоліків: | Вплив на набір даних: |
| - відсутність відповіді від одиниці | Весь вектор даних залишається порожнім (або всі елементи відхилені) |
| - відсутність відповіді від елементу | Один з кількох елементів є порожнім або відхиляється |
| - відсутність відповіді від одиниці/часткова відсутність відповіді | Всі дані з одного або кількох кластерних елементів відсутні, наприклад, члени однієї або кількох сімей відмовляються, дані отримуються від інших |

**Відсутність відповіді від одиниці**

Всі організації що проводять дослідження, відстежують величину і розподіл відповіді або відсутності відповіді. Однак, незважаючи на зусилля, стандартизовані способи розрахунку цих категорій відсутні (див., наприклад, Lynn та ін., 2003).

Як правило, неотримання відповіді при дослідженнях домогосподарств, класифікується за такими категоріями: Відсутність контакту, відмова та інші причини. Крім того, у разі можливості слід розрізняти помилку репрезентативності.

В бізнес-дослідженнях (які часто є обов’язковими) питання охоплення мають тенденцію бути більш впливовими при дослідженнях домашніх господарств. У деяких випадках підприємства відмовляються відповідати, тому що вони можуть відчувати, що їхні конкуренти можуть якось отримати стратегічної ринкову інформацію. Органи статистики прагнуть до того, щоб більш домінуючі підприємства надавали відповідь в будь-якому випадку для забезпечення довіри до результатів.

## Повторне зважування з метою коригування відсутності відповіді від одиниці

Якщо розподілити осіб, які дають і які не дають відповідь випадковим чином по всій вибірці, відсутність відповіді може бути ідентифікованою як *проігноровано,* в цьому випадку можна застосувати просту модель τ *= m /n ,* за якою розмір вибірки замінюється кількістю осіб, що дали відповідь при зважуванні. В цьому випадку єдиним недоліком буде збільшена дисперсія вибірки. На жаль, така проста модель, як правило, не діє, і дані можуть бути спотворені за рахунок ефекту відсутності відповіді, *не проігнорована відсутність відповіді*. Найбільш часто вживані методи повторного зважування є постстратіфікацію та оцінкою у вигляді співвідношення. Нижче наводимо деякі інші методи.

**Пост-стратифікація**

Ми можемо легко здійснити *пост-стратифікацію* щодо нашого набору даних (див. Розділ 3.4). При коригуванні відсутності відповіді найкращим способом є знаходження субгруп, в яких рівні отримання відповіді/відсутності відповіді максимально відрізняються. Ідея полягає в тому, щоб знайти однорідні підгрупи, які виявляють поведінку відповіді. Для цієї задачі потрібна значна кількість таблиць, після створення яких зважування перетворюється на відносно просту процедуру.

**Оцінка у вигляді співвідношення**

*Оцінку у вигляді співвідношення* (див. 3.4) також можна використовувати після збору даних безпосередньо для оцінки змінної дослідження *y*, або шляхом зважування. Нам потрібні підсумки щодо безперервних змінних на рівні сукупності або на рівні області, а також відповідні спостереження *xk* для кожного респонденту.

Оцінка у вигляді співвідношення дає хороші результати для змінної дослідження *y*  тільки у випадку сильної позитивної кореляції *x* та *y.* Зверніть увагу, що оцінка у вигляді співвідношення має систематичну помилку за визначенням, якщо лінія істинної регресії не проходить через вихідну точку.

**Емпіричні однорідні групи відповіді (RHG)**

Емпіричні *однорідні групи відповіді* (RHG), які часто називають класи зважування, означають, що елементи вибірки діляться на взаємно прості підгрупи відповідності ймовірностей відповідей. Технічно це подібно до пост-стратифікації, але діє тільки для інформації на рівні вибірки, - інформації, що не відноситься до сукупності.

Уявіть, що при структурі SRS-WOR частину вибірки, що дає відповідь, можна поділити відповідно до RHG на *L*  категорій:



Таким чином, остаточна змінена вага дорівнюватиме *wk = ak* /τ*l* , *k ϵ Gl* , де *ak*  є вагами структури, а *Gl*  є однорідною групою відповіді *l*.

Отримання належної оцінки дисперсії без зміщення для оцінки класу зважування є дуже трудомісткою процедурою і, тому ми її тут не наводимо.

**Моделювання ймовірності явної відповіді**

Моделювання є продовженням табличних даних. Таким чином моделі можуть підходити або на індивідуальному рівні, або на частотному рівні. Ідея знову полягає в тому, щоб поділити вибірку на окремі групи за ймовірністю відповідей. Проте, модель використовується для того, щоб допомогти цій задачі, а також, щоб оцінити ймовірності відповідей на основі моделей. Моделі логістичної регресії (див. 3.4) є найбільш популярними в цьому випадку у зв'язку з відповідним характером явища.

Гарна модель враховує основні ефекти багатьох змінних і, при необхідності, відповідні взаємодії. В результаті прогнозовані значення є більш гладкими, ніж оригінальні емпіричні ймовірності відповідей.

**Класифікація ваг**

*Оцінка класифікації*, ймовірно, є найбільш використовуваним методом повторного зважування в даний час. Він є універсальним і поєднує в собі гарні властивості моделювання відсутності відповідей, з необхідністю узгодженості з допоміжною інформацією, тобто оцінки вибірки щодо обраного розподілу допоміжної інформації відповідають розподілу сукупності. Також наявні наближені варіанти дисперсії вибірки (див., наприклад, Deville & Särndal, 1992; Särndal & Lundström 2005).

## Підстановки для коригування відсутності відповіді від елементу

Підстановки є основним методом для компенсації відсутності відповіді від елементу. Він також застосовується для корекції помилкових значень, знайдених при перевірці даних. Тому цю стадію робіт часто називають «редагування та підстановки». Основні методи - це логічні підстановки, підстановки на основі реального донора та модельні підстановки. В Європейських дослідницьких рамкових програмах 4 і 5 містяться проекти, що називаються AUTIMP і EUREDIT, які забезпечили багато інструментів для бюро статистики щодо редагування та підстановок.

**Логічні підстановки** пов’язані з редагуванням даних. В даний час вона виконує роль програми з введення даних: програми інтерв’ю за допомогою комп’ютера, програми отримання даних з адміністративних джерел і різні програми для введення даних для бізнес-спостережень оснащені набором логічних процедур перевірки, які можуть виявляти помилки, а також надавати правильні значення або розрахункові значення «краще припущення» для відсутніх або неправильних даних.

**Підстановки на основі реального донора** спирається на відповіді від однієї одиниці або подібної одиниці. В бізнес-дослідження вже стало традицією підставляти відсутні значення значеннями з попереднього вимірювання одного щодо тієї ж одиниці. Цей метод називається методом *холодних підстановок*. Іноді попередні значення помножуються на середнє значення зміни для змінної, що розраховується для тієї ж галузі.

Якщо значення запозичується від сусідньої одиниці (фізично це наступне або попереднє спостереження) *метод гарячих підстановок*. Традиційні гарячі підстановки часом призводять до неправдоподібних ситуацій і тому вони поліпшуються за допомогою *методу ближніх підстановок*, в якому підбираються однакові одиниці за допомогою певних параметрів або шляхом відбору субсукупностей.

Значення може безпосередньо отримуватися з найближчої одиниці або на підставі лотереї з пулу подібних одиниць.

**Підстановки з моделлю-донором** означають ситуацію, коли використовується певна статистична модель для оцінки відсутнього (або неправильного) значення. Як правило, це також відбувається в бізнес-дослідженнях, в яких можна використовувати прийнятні відповіді, які б задовольняли моделі, а потім вони застосовуються до одиниць з відсутніми значеннями. Найпростіший спосіб полягає в використанні загального середнього значення або відповідного значення області. Підстановка середніх значень не може вважатися задовільною через тенденцію значного зниження дисперсії у випадку наявності значної кількості відсутніх значень. Підстановки на основі регресії або іншої статистичної моделі дають набагато кращі результати в цьому відношенні. Можна також зменшити тенденцію зниження дисперсії, додавши певну додаткову дисперсію з відповідного розподілу, що дасть *стохастичні підстановки*, а не *детерміновані підстановки*.

В принципі всі вищевказані методі мають справу з *однією підстановкою*, тобто відсутнє значення підставляється тільки одним значенням. *Множинні підстановки* - це інший підхід, який дозволяє вирішити багато проблем шляхом підстановок одного відсутнього значення цілим рядом нових значень (див. Rubin 1987 або Schafer, 1997). Це статистично обґрунтований метод, але бюро статистики (і часто їх клієнти) ще не запровадили ці методи у своїх програмах роботи.

# Програмне забезпечення

Дві великомасштабні статистичні програми, SAS і SPSS, містять конкретні модулі відбору зразків та аналізу даних дослідження. В програмі SAS ці процедури включені в модуль STAT, починаючи з версії 8 і вище. У SPSS є додатковий модуль Complex Samples, який має придбаватися окремо, починаючи з версії 11 і вище.

Мова програмування R широко використовується в наукових дослідженнях і деякі автори надали код для вибіркових досліджень та аналізу. Їх можна знайти на різних веб-сторінках, що містять програмний код R, наприклад на [www.r-project.org/.](http://www.r-project.org/)

Програмне забезпечення для вибірки

SAS/Stat версія 8 і вище Proc SurveySelect (www.sas.com)

SPSS/Complex Samples версія 11 і вище CSPLAN разом з CSSELECT

(www.spss.com)

Програмне забезпечення для зважування

|  |  |  |
| --- | --- | --- |
| Додаткові програми до SAS: | CALMAR2 | (INSEE/Франція) |
|  | CLAN97 | (Управління статистики Швеції) |
| Додаткові програми до SPSS: | g-CALIB | (Управління статистики Бельгії) |
| BLAISE компонент | BASCULA | (Управління статистики Нідерландів) |

Програмне забезпечення для редагування та підстановок

SAS/Stat версія 8 і вище Proc ІМ та MIANALYZE (множинні підстановки)

Додаткові програми до SAS: BANFF/GEIS (Управління статистики Канади)

 CONCORD, DIESIS,

QUIS (ISTAT/Італія) IVEWARE

([www.isr.umich.edu/src/smp/ive/)](http://www.isr.umich.edu/src/smp/ive/%29)

Програмне забезпечення для «стандартних» оцінок

SAS/Stat версія. 9 і вище Proc Surveymeans Proc Surveyfreq

Додаткові програми до SASs: CLAN97 (Управління статистики

 Швеції)

 POULPE (INSEE/France)

 GES (Управління статистики

 Канади)

SPSS версія 11 і вище CSDESCRIPTIVES

Complex Samples CSTABULATE

Додаткові програми до SPSS: g-CALIB (Управління статистики

 Бельгії)

Компонент BLAISE BASCULA (Управління статистики

 Нідерландів)

SUDAAN (www.rti.org)

Програмне забезпечення для аналітичних цілей

SAS/Stat версія 9 і вище Proc Surveyreg

 Proc Surveylogistic

(GLM, MIXED, NLMIXED,…)

SPSS версія 11 і вище CSGLM, CSORDINAL,

Complex Samples CSLOGISTIC SUDAAN

WesVAR (www.westat.com)

# Література

Andersson, C. & Nordberg, L. (1994). A Method for variance estimation of non-linear func- tions of totals in surveys ! theory and a software implementation. *Journal of Official Statis- tics*, Vol. 10, No. 4, 395-406.

Andersson C. (2002). CLAN97 v. 3.1. Supplement to A User’s Guide to CLAN97. (Unpub- lished).

Biemer P. & Lyberg L. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons.

Brewer K. & Hanif F. (1983). *Sampling with Unequal Probabilities*. New York: Springer- Verlag.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons.

Deville J-C., & Särndal C-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, Vol. 87, No. 418, 376382.

Deville J-C., Särndal C-E., & Sautory O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, Vol. 88, No. 423, 1013-1020.

European Union (1998). *Council regulation (EC) No 577/98 of 9 March 1998 on the organi- sation of a labour force sample survey in the Community*.

Groves R.M., Dillman D.A., Eltinge J.L., & Little R.J.A. (eds.). (2002). *Survey Nonresponse*. New York: John Wiley & Sons.

Groves R.M., Fowler F.J., Couper M.P., Lepkowski J.M., Singer E. & Tourangeau R. (2004).

*Survey Methodology*. Hoboken: John Wiley & Sons.

Hansen M.H., Hurvitz W.N. & Madow W.G. (1953). *Sample survey methods and theory*. New York: John Wiley & Sons.

Kish L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Lehtonen R. & Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. 2nd ed. Chichester: John Wiley & Sons.

Lehtonen R., Särndal C.-E. & Veijanen A. (2005). Does the model matter? Comparing model- assisted and model-dependent estimators of class frequencies for domains. *Statistics in Tran- sition*, 7, 649–673.

Lessler J. & Kalsbeek W. (eds.). 1992. *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Lohr S. (1999). *Sampling:* *Design and analysis*. Pacific Grove: Duxbury Press.

Lyberg L., Biemer P., Collins M., de Leeuw E. D., Dippo C., Schwarz N. & Trewin D. (eds.). (1997). *Survey Measurement and Process Quality.* New York: John Wiley & Sons.

Lynn P., Beerten R., Laiho J. & Martin J. (2003). Towards standardisation of survey outcome categories and response rate calculations. *Research in Official Statistics,* Vol. 5 1/2002, 61-84*.*

Oh J.L. & Scheuren F. (1983). Weighting Adjustment for Unit Nonresponse, in W.G. Madow, I. Olkin & D.B. Rubin (eds.) *Incomplete Data in Sample Surveys*, Vol 2., 143-184. New York: Academic Press.

Rubin D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Sautory O. (1993). La macro CALMAR. Redressement d'un échantillon par calage sur mar- ges. I.N.S.E.E., Série des documents de travail nЕ F 9310. Paris: I.N.S.E.E.

Sautory, O. & Le Guennec, J. (2005). La macro CALMAR2. Redressement d'un échantillon par calage sur marges. I.N.S.E.E. Paris: I.N.S.E.E.

Schafer J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall. Smith T.M.F. (1990). Post-stratification. *The Statistician*, 40, 315 ! 323.

Sundgren B. (1999). Information systems architecture for national and international statistical

offices. Guidelines and recommendations. *Conference of European Statisticians.* *Statistical Standards and Studies.* No 51. Geneva: United Nations.

Särndal C-E. & Lundström S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: John Wiley & Sons.

Särndal C-E., Swensson B. & Wretman J. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

United Nations (1950). The Preparation of Sampling Survey Reports. *Statistical Papers* Ser. C No. 1 (Revised). New York.

United Nations (1964). Recommendation for the Preparation of Sample Survey Reports (Pro- visional Issue). *Statistical Papers* Ser. C No. 1 Rev.2. New York.

Wolter K. (2007). *Introduction to Variance Estimation,* Second Edition. New York: Springer.

# Посилання на веб-ресурси

AUTIMP (програмне забезпечення для автоматичних підстановок для бізнес-спостережень та переписів населення) див. EUREDIT

EUREDIT (Розробка і оцінка нових методів для редагування та підстановок): <http://www.cs.york.ac.uk/euredit/>

Сайт з забезпечення якості Євростату: <http://ec.europa.eu/eurostat/quality>

Кодекс норм Європейської статистики Стандарти якості ЄСС:

Стандартний доповідь з якості

Стандартні характеристики якості

Як підготувати доповідь з якості

Довідник з поліпшення якості шляхом аналізу технологічних параметрів DESAP - самостійна оцінка для керівників, що зайняті в дослідженні

Міжнародний валютний фонд. Система оцінки якості даних. 2003, <http://dsbb.imf.org/vgn/images/pdfs/dqrs_Genframework.pdf>

Організація економічного співробітництва та розвитку. База даних OECD PISA, 2001: <http://pisaweb.acer.edu.au/oecd/>

Організація економічного співробітництва та розвитку. Система забезпечення якості та рекомендації щодо статистичних даних ОЕСР. 2003. <http://www.oecd.org/dataoecd/26/42/21688835.pdf>

Управління статистки Фінляндії: Рекомендації щодо забезпечення якості для офіційної статистики:: <http://stat.fi/tk/tt/laatuatilastoissa/alku_en.html>

Віртуальна лабораторія VLISS у вибірковому дослідженні:

[http://www.math.helsinki.fi/VLISS/.](http://www.math.helsinki.fi/VLISS/)

**Додаток 1**. Приклад вибірки з використанням вибірки Bernoulli

Ми створюємо основу вибірки, що складається 2000 елементів і нам потрібно для вибірки обрати близько 200 одиниць. Застосовуються такі ж самі символи, що і на сторінці 15.

Доля вибірки UNI = 200/2000 = 0.1. Всім елементам в основі вибірки присвоюється псевдо випадкове число з рівномірного розподілу, PI. Обираються такі елементи, для яких PI ≤ 0.1, а індикаторі вибору дається значення 1. Якщо одиниця не вибрана, IND присвоюється 0.

Програмний код для SAS є дуже простим: Данні Bernoulli;

UNI=200/2000; /\* Встановлюються межі відбору \*/

Do I=1 to 2000; /\* Створюється основа вибірки з 2000 елементів\*/ PI=Ranuni(0); /\* Кожному елементу присвоюється випадкове число\*/ If PI<=UNI then IND=1; /\* Здійснюється перевірка, чи була включена одиниця \*/

Else IND=0; /\* або ні \*/ Результат;

End;

Proc Print; Sum IND;

Run;

|  |  |  |  |
| --- | --- | --- | --- |
| I | UNI | PI | IND |
| 1 | 0.1 | 0.83976 | 0 |
| 2 | 0.1 | 0.50375 | 0 |
| 3 | 0.1 | 0.08013 | 1 |
| 4 | 0.1 | 0.87756 | 0 |
| 5 | 0.1 | 0.13501 | 0 |
| 6 | 0.1 | 0.41416 | 0 |
| 7 | 0.1 | 0.10639 | 0 |
| 8 | 0.1 | 0.28283 | 0 |
| 9 | 0.1 | 0.16496 | 0 |
| 10 | 0.1 | 0.88332 | 0 |
| ... |  |  |  |
| 1991 | 0.1 | 0.67351 | 0 |
| 1992 | 0.1 | 0.11558 | 0 |
| 1993 | 0.1 | 0.78235 | 0 |
| 1994 | 0.1 | 0.66004 | 0 |
| 1995 | 0.1 | 0.08314 | 1 |
| 1996 | 0.1 | 0.19041 | 0 |
| 1997 | 0.1 | 0.77828 | 0 |
| 1998 | 0.1 | 0.07666 | 1 |
| 1999 | 0.1 | 0.53644 | 0 |
| 2000 | 0.1 | 0.35678 | 0 |
| Sum |  |  | 201 |

У вибірці Bernoulli розмір вибірки є випадковою величиною і цей приклад показує, що ми отримали занадто велику одну одиницю. Найпростішим способом отримати фіксований розмір вибірки полягає у сортуванні основи вибірки за випадковим числом і обранні рівно 200 випадків (з початку, з кінця або з будь-якої точки поки для відбору використовуються випадкові числа).

Європейська Комісія

**Довідкові рекомендації вибірки дослідження - Введення у структуру вибірки та методи оцінки**

Люксембург: Бюро офіційних публікацій Європейських Спільнот.

2008 — 36 стор. — 21 x 29,7 см.

ISBN 978-92-79-07840-8

ISSN 1977-0375