Сальнікова Світлана Анатоліївна,
асистент кафедри соціології Волинського державного університету імені Лесі Українки

Алгоритм систематичного відбору з випадковим початком: наукове обгрунтування та практична доцільність

В статье рассматриваются простые вероятностные методы отбора: собственно случайный и систематический, подробно рассмотрен этап, связанный с отбором единиц. Автором предлагается математическая модель систематического отбора со случайным началом, а также компьютерная программа для реализации частичного случая данного метода. Описывается алгоритм случайного отбора, который модифицирует простой отбор с фиксированным шагом. Для подтверждения эффективности предложенного алгоритма проведено тестирование применения программы.

The article represents the simple probabilistic methods of selection, namely actually accidental and systematic; the stage connected with the selection of units is being considered in details. The scientific novelty of the given work consists in the creation of a mathematical model of the systematic selection with the accidental beginning and a computer program for realizing the partial case of the given method. The aim of the article proper is the creation of an algorithm of the accidental selection, modifying simple selection with the fixed step and confirmation of the scientific and practical expediency of the method. For the purpose of confirming the suggested algorithm efficiency, the testing of the programme results of two independent experiments by means of the criterion χ^2 has been conducted.

Теорія вибіркового методу – важливий розділ математичної статистики, її твердження базуються на поняттях теорії імовірності – випадкових подіях і випадкових величинах, а діапазон застосування вибіркового методу досить широкий. Останнім часом досить поширеними стали опитування громадської думки та передвиборні опитування [1, с.18], саме вони зіграли велику роль у ознайомленні громадськості з методами вибіркового дослідження. Його бузумовні переваги: він дешевший, більш достовірний, дозволяє зменшити час, відведений на опитування, має ширшу область застосування [2, с.15].

Власне темою статті є алгоритм випадкового відбору, що модифікує простий відбір із фіксованим кроком, а з*авданням даної роботи* є

- сформулювати (означити) поняття вибірки,
- представити етап вибіркового дослідження, пов'язаний з відбором одиниць,
- подати математичну модель систематичного відбору з випадковим початком та створити комп'ютерну програму для реалізації цього методу,
- провести тестування програми (оцінити ефективність запропонованого алгоритму) за допомогою критерію χ^2.

Тема вибіркового методу є невід'ємною частиною будь-якого методичного посібника, присвяченого технології дослідження. Серед авторів слід відзначити таких як Паніна Н.В. [3, с.63], Батигін Г.С. [4, с.137], Єрмолаєв А. [5,], Г. Шварц [6] та інші.

Однією із задач, які постають перед соціологом при проведенні дослідження, є вміння створити вибірку, адже репрезентативність інформації в емпіричному дослідженні забезпечується в процесі організації вибірки. Поняття вибірки в соціології (статистиці, маркетингу) розглядається у двох значеннях. По-перше, вибірка – це елементи генеральної сукупності, які підлягають вивченню, тобто вибіркова сукупність[1]. По-друге, вибірка – це метод (або процедура, процес) формування вибіркової сукупності при необхідній умові забезпечення репрезентативності. Виділяють різні типи вибірки (добору) і види вибірок[2]. Але, якщо виходити з основних принципів підходу до добору одиниць з генеральної сукупності, то існує три основних підходів (інші – це ніщо інше, як різноманітні варіанти): простий імовірнісний, квотний та стихійний.

Прості імовірнісні методи є як окремими (самостійними) методами випадкового відбору, так і методами, які застосовують як допоміжні при застосуванні інших методів.

Прості імовірнісні вибірки формують при наявності повних списків елементів генеральної сукупності. Існує кілька загальновідомих методів створення власне випадкової вибірки:

1. За допомогою *таблиці випадкових чисел*.
2. За допомогою *генератора випадкових чисел*.
3. За допомогою *схеми урн*.

При створенні вибірки даними методами можна легко уникнути систематичної помилки внаслідок повного дотримування принципу випадковості. Недоліки цих методів можна представити двома пунктами:

---

[1] При цьому кожен елемент досліджуваної сукупності має рівну імовірність потрапити у вибірку. Така властивість називається принципом випадковості.

[2] Методи випадкового відбору у вигляді схеми подав А. Єрмолаєв [5].

- Обов’язково потрібно мати увесь список елементів генеральної сукупності, тобто списки людей з необхідними даними, наприклад, з адресами й телефонами.
- Для сукупності великого об’єму дані методи не застосовують у зв’язку з громіздкими підготовчими роботами. Складність проявляється і при опитуванні, адже ми отримуємо точні номери, а тому, конкретні імена осіб, яких слід опитати.

Отже, методи створення власне випадкової вибірки підходять лише для невеликих об’ємів генеральної сукупності. Але соціологічні дослідження, як правило, є дуже об’ємними. Тому для них краще застосовувати *систематичний відбір з випадковим початком*.

Нехай генеральна сукупність є деякою впорядкованою множиною, а доля вибірки становить $\frac{n}{N} = f$. Тоді число $c = \frac{1}{f} = \frac{N}{n}$ називається інтервалом або кроком систематичного відбору. Тобто з кожних c одиниць сукупності буде обрано один. Для цього випадковим методом з проміжку від 1 до c вибирають деяке число i, яке й буде випадковим початком. До вибірки ввійдуть такі одиниці сукупності:

$$i, [i + c], [i + 2 \cdot c], ..., [i + (n - 2) \cdot c], [i + (n - 1) \cdot c].$$

Приклад 1. Нехай f=0,25, тоді c=4, і випадковим методом з чисел 1, 2, 3, 4 було обрано число 3. Отже, сукупність поділимо на інтервали по чотири одиниці в кожному, і вибирати будемо з інтервалу 3-ій елемент. Матимемо вибірку з таких елементів: 3, 7, 11, 15, 19, ...

Даний приклад підходить тоді, коли $N \vdots c$, тобто коли сукупність ділиться на інтервали з однаковою кількістю одиниць або коли c є числом цілим. Він є спрощеним варіантом систематичного відбору. Продемонструємо інший, більш загальний приклад.

Приклад 2. Нехай N=3000, n=700, тоді крок відліку c=4,3. Створення вибірки проведемо в три етапи.

1) З чисел 1, 2, 3, 4 випадково обране число буде першим у новому списку, нехай це число 3, елементи під номерами 1, 2 слід перенести в кінець списку, матимемо: 3, 4, 5, 6, …, 2999, 3000,1,2.

2) Оскільки об’єм сукупності N не ділиться на потрібну нам кількість елементів вибірки n без остачі, проведемо наступні міркування:

$$\frac{3000}{700} = \frac{30}{7} \cdot \frac{100}{100} = \frac{7 \cdot 4 + 2}{7} \cdot \frac{100}{100},$$

звідки бачимо: сукупність слід поділити на (7-2)*100=500 інтервалів по 4 елементи і 2*100=200 інтервалів по (4+1)=5 елементів, тобто на кожні 7 інтервалів припадає 5

інтервалів по 4 елементи і 2 – по 5 елементів. Вибравши випадково два числа з проміжку 1, …,7, знатимемо номери інтервалів, які матимуть по 5 елементів. Наприклад, це числа 1 і 6, тоді п'ятиелементними будуть 1, 6, 8, 13, 15, 20, … інтервали.

3) І нарешті, знайдемо номер елемента з кожного інтервалу, який входитиме до вибірки: це випадково обране число з проміжку від 1 до 4.

Узагальнимо даний метод: Нехай N – впорядкована множина елементів генеральної сукупності, n – кількість елементів вибірки, c – крок відліку. І нехай c не є цілим числом[3], тобто N ділиться на n з остачею. Аналогічно до прикладу 2 розіб'ємо створення вибірки на три етапи:

1) З проміжку [1, c] випадково оберемо перший елемент нового списку. Нехай це буде елемент i, тоді матимемо ту ж сукупність, але зміщену вліво на $(i-1)$ елемент:

$$i, i+1, \ldots, c, c+1, \ldots, N, 1, 2, \ldots, i-1 .$$

2) Оскільки N не ділиться на n без остачі, то визначимо: скільки буде інтервалів з (c) елементами, а скільки з (c+1), і як вони будуть розподілені по сукупності. Для цього спочатку із співвідношення $\frac{N}{n}$ винесемо спільний множник виду 10^k (якщо такий є)[4], а потім розкладемо ділене число N за дільником n:

$$\frac{N}{n} = \frac{10^k}{10^k} \cdot \frac{N_1}{n_1} = \frac{10^k}{10^k} \cdot \frac{n_1 \cdot c + r}{n_1}$$

З формули видно: сукупність слід поділити на $(n_1 - r) \cdot 10^k$ інтервалів з (c) елементами і на $r \cdot 10^k$ інтервалів з (c+1) елементом таким чином, щоб на кожні n_1 інтервалів припадало $(n_1 - r)$ інтервалів з (c) елементами і r інтервалів з (c+1).

Щоб визначити, які саме інтервали матимуть (c+1) елемент, потрібно з проміжку 1, 2, …, n_1 обрати r випадкових чисел, які будуть номерами шуканих інтервалів. Нехай $\underbrace{r' \quad r'' \quad \ldots \quad r^{(k)}}_{r}$, тоді кожен

[3] В якості кроку c з проміжку [1,c] обирають просте число або число, що не є дільником розміру генеральної сукупності [7, с.22].
[4] Див. Зауваження 4

$$\underbrace{r',r'',\ldots,r^{(k)},r'+n_1,r''+n_1,\ldots,r^{(k)}+n_1,r'+2\cdot n_1,r''+2\cdot n_1,\ldots,r^{(k)}+2\cdot n_1,\ldots,r'+c\cdot n_1\cdot 10^k,r''+c\cdot n_1\cdot 10^k,\ldots,r^{(k)}+c\cdot n_1\cdot 10^k}_{r\cdot 10^k}$$

інтервал матиме (c+1) елемент.

3) Після поділу сукупності на інтервали залишилось знайти номер елемента з проміжку [1,c], який попаде у вибірку. Якщо випадково обраним є число i, то з кожного інтервалу i-ий елемент буде елементом вибірки.

Зауваження 1: Кількість інтервалів, на які треба розбити сукупність N, рівна кількості елементів вибірки n.

Це пояснюється тим, що з кожного інтервалу береться лише один елемент.

Зауваження 2: При поділі сукупності на інтервали має місце нерівність $r < n_1$.

В розкладі числа $N_1 = n_1 \cdot c + r$ число r є остачею від ділення N_1 на n_1, а за властивістю ділення остача не перевищує дільника (n_1). Очевидно, що рівність виконуватиметься тоді, коли N ділиться на n без остачі.

Зауваження 3: Кожна сукупність фактично складається з 10^k блоків, в кожному блоці n_1 кількість інтервалів, з яких $(n_1 - r)$ інтервалів з (c) елементами і r інтервалів з (c+1) елементами. При цьому добуток блоків на інтервали є нічим іншим, як об'ємом вибірки $10^k \times n_1 = n$.

Зауваження 4: В розкладі (2) важливу роль відіграє множник 10^k. Тут слід виділити такі випадки:

1. $k = 0$, тобто множник 10^k відсутній.
2. $k = 1$, тобто обсяги генеральної N та вибіркової n сукупностей кратні 10.
3. $k \geq 2$, тобто $10^k = 100,1000,\ldots$.

Випадки 1 та 2 вимагають більш громіздкого розгляду та глибоких математичних знань, вони є перспективою для подальшої роботи в даному напрямку. Якщо ж $k \geq 2$, то особливих труднощів в побудові інтервалів не виникає. Цей випадок є основою написання програми для знаходження номерів тих елементів, яких слід опитати на підставі, метода застосування систематичного відбору з випадковим початком.

Програма має такий зовнішній вигляд:

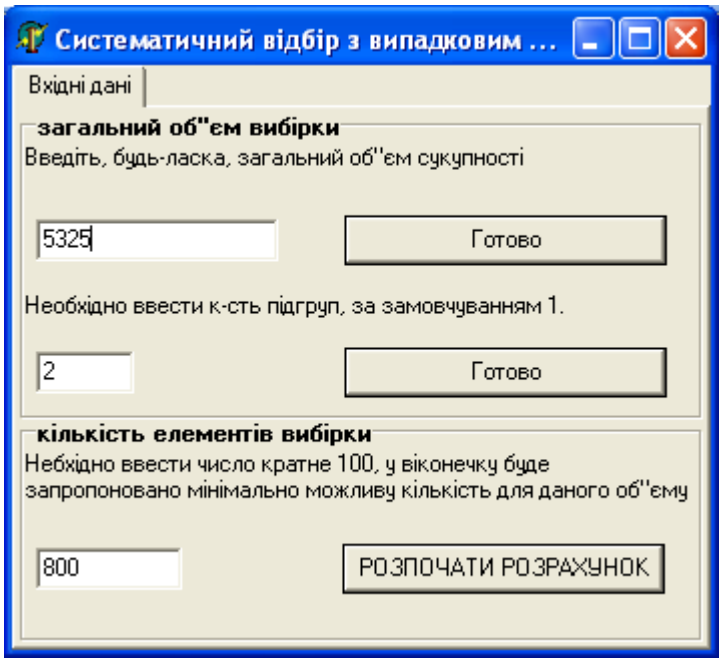

Рисунок 1.

Загальний обсяг сукупності не обов’язково має бути кратним 100. При натисканні клавіші «Готово» або клавіші табуляції програма автоматично зменшить число до найменшого кратного 100, тобто матимемо 5300. З отриманим числом легше будувати та розподіляти інтервали. Проте 25 відкинутих елементів насправді нікуди не зникають. При відборі одиниць з кожного інтервалу верхньою межею є число, початково введене користувачем, тобто 5325. Порядок виконання алгоритму такий: спочатку поділ на інтервали, потім відбір елементів з кожного інтервалу, потім до кожного з відібраних елементів додаємо число, що є випадковим початком. Останнє отримане число порівнюється з верхньою межею, процедура додавання повторюється, доки не вичерпаються усі числа з проміжку від 1 до 5325.

Після нажатої клавіші «Розпочати розрахунок», пропонується користувачу самому обирати випадкові числа:

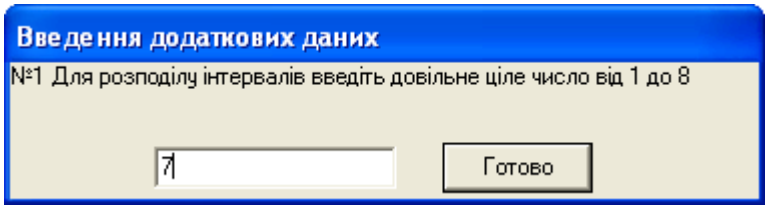

Рисунок 2.

В результаті обчислень програми матимемо:

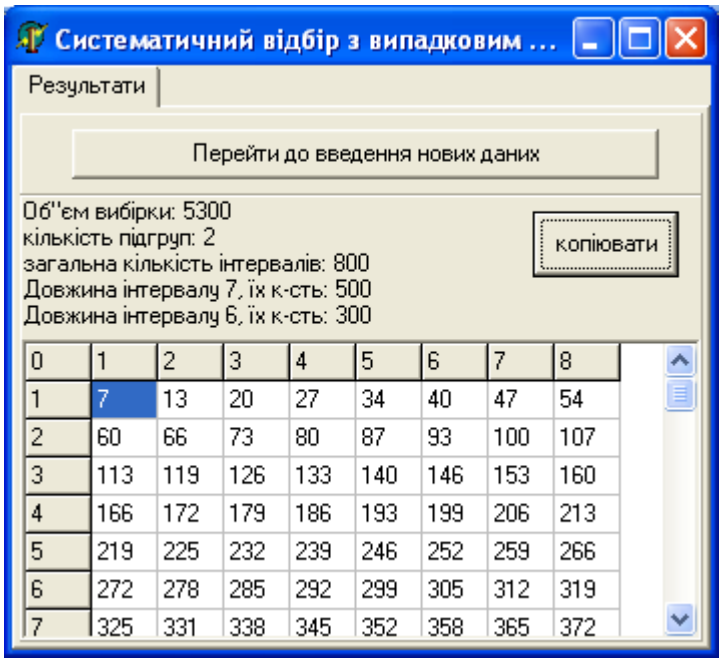

Рисунок 3.

Кнопка «копіювати» дозволяє експортувати дані в програми Word або Excel, а «Перейти до введення нових даних» провести інші розрахунки.

Дана програма є процедурою отримання рівномірно розподіленої випадкової сукупності[5] номерів (тобто свого роду датчиком рівномірно розподілених випадкових чисел[6]). А будь-яку послідовність випадкових чисел, перш ніж застосовувати, слід детально перевірити. Для тестування застосуємо критерій χ^2.

Було проведено два незалежних експерименти: з двох сукупностей різного об'єму побудовано вибірки різними способами (причому інтервал відбору однієї c=3, а іншої c=8); в таблиці наведено частоти попадання чисел в інтервали однакової довжини. Поряд із значеннями емпіричних частот N_{ij} подано в дужках теоретичні значення, обчислені за формулою $N_{ij} = \frac{1}{N} \cdot N(x_i) \cdot N(y_j)$.

Таблиця 1. ***Результати двох незалежних експериментів***

X	Y					N(X)
	y_1	y_2	y_3	y_4	y_5	
x_1	209 (213)	210 (213)	208 (204)	211 (214)	209 (208)	$N(x_1)$=1047
x_2	126 (122)	125 (122)	112 (116)	115 (118)	118 (119)	$N(x_2)$=596
N(Y)	$N(y_1)$=335	$N(y_2)$=335	$N(y_3)$=320	$N(y_4)$=326	$N(y_5)$=327	N=1643

[5] Випадкова сукупність називається рівномірно розподіленою

[6] Процедурним аспектам побудови датчиків випадкових чисел присвячена перша половина другого тому монографії Д. Кнута [7].

Обчисливши статистику критерію χ^2 за загальною формулою $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(N_{ij} - N_{ij}^0)^2}{N_{ij}^0}$ отримаємо невелике значення χ^2=0,67, при цьому число ступенів свободи $f = (k-1)(l-1) = 4$. Табличне значення [8, с.255], що відповідає p=0,95 і $f = 4$, є таким $\chi_0^2 = 9{,}49$, тобто нерівність $\chi^2 < \chi_0^2$ є суттєвою, а це говорить про те, що відхилення даних теоретичної таблиці від емпіричної має випадковий характер, тобто отримані в результаті експериментів послідовності є випадковими.

Дана програма має і практичну доцільність. Наприклад, соціологічні дослідження серед молоді, що навчається. В будь-якому регіоні кількість училищ та технікумів переважає над кількістю університетів та інститутів, причому чисельність студентів в перших (I) значно нижча, ніж у вищих навчальних закладах (II) . Розподіливши кількість номерів вибірки серед навчальних закладів методом квот і маючи повні списки студентів (вони є завжди, причому в електронному вигляді), до груп I та II можна застосувати систематичний відбір з випадковим початком. Списки слід експортувати в програму Excel у вигляді [№ (номер по списку групи), П.І.Б., назва закладу, № групи][7] і занумерувати числами натурального ряду. При створенні вибірки немає проблем з повними списками елементів генеральної сукупності, і менш вагомими стають недоліки, що виникають в процесі опитування.

Література:

1. Опитування громадської думки / За ред. Н.В. Паніної. – К.: Інститут соціології НАН України, 2003. – 80 с.

2. Кокрен У. Методы выборочного исследования. – М., 1976.

3. Паніна Н.В. Технологія соціологічного дослідження. – К.: «Наукова думка», 1996. – 232 с.

4. Батыгин Г.С. Обоснование научного вывода в прикладной социологии. М.: Наука, 1986 – 271 с.

5. Ермолаев А. Выборочный метод в социологии. – М., 2000

6. Г. Шварц. Выборочный метод: Пер. с нем. – М.: «Статистика», 1978. – 213 с.

7. Д. Кнут. Искусство программирования для ЭВМ. т.2. Получисленные алгоритмы. Пер. с англ. - М.: «Мир», 1977 – 728 с.

8. Паніотто В.І., Максименко В.С., Харченко Н.М. Статистичний аналіз соціологічних даних. – К.: Вид. дім «КМ Академія», 2004. – 270 с.: іл..

[7] ідеальним варіантом було б створення файлу із записами в такому вигляді, щоб програма сама зчитувала і видавала не номери, а записи.