УДК 330.4:519.22:004.9
DOI 10.31654/2663-4902-2025-PP-1-121-129

Фетісов В. С.
кандидат економічних наук, доцент,
доцент кафедри інформаційних технологій, фізико-математичних та економічних наук
Ніжинського державного університету імені Миколи Гоголя
fetisov.vs@ndu.edu.ua
orcid.org/0000-0003-4316-9060

ПРАКТИЧНІ ПИТАННЯ ФОРМУВАННЯ ВИБІРКОВИХ СУКУПНОСТЕЙ У STATISTICA

Сучасний світ – це світ інформації. Завдяки використанню Інтернет користувачі отримали доступ до величезних обсягів найрізноманітніших даних. Але ці дані стають по справжньому корисними тільки після того, як виявляються існуючі в них закономірності, що дозволяє надалі застосовувати їх у практичній діяльності.

Зрозуміло, що розуміння сутності даних можна розкрити тільки після проведення їх статистичного аналізу. Але аналіз великих обсягів даних вимагає застосування обчислювальної техніки. При цьому використання програмного забезпечення загального призначення на зразок Excel не завжди дозволяє вирішити конкретне завдання статистичного аналізу. У цьому разі дослідники мають вміти користуватися спеціалізованим програмним забезпеченням – статистичними пакетами обробки даних, серед яких є два беззаперечних лідери – STATISTICA та SPSS. У статті розглядається робота із STATISTICA. Використання пакету STATISTICA може суттєво прискорити, а – головне – зробити більш якісним статистичний аналіз даних.

Не завжди у дослідника є можливість працювати з усією сукупністю даних, через що доводиться використовувати частину даних, тобто вибіркову сукупність, на підставі якої і робляться надалі висновки відносно усієї (генеральної) сукупності в цілому. Досить часто взагалі досліднику потрібно відібрати тільки певну частину даних з наявних даних. Питання роботи з вибірковою сукупністю у статистиці розглядаються в окремому розділі: вибірковому методі, а способи формування вибіркових сукупностей є одним з підрозділів вибіркового методу.

Під час проведеного дослідження розглянуто способи формування як шляхом створення підмножин, так і за допомогою механізму створення випадкової вибірки за різними варіантами з існуючих робочих таблиць даних. При цьому наведені приклади застосування з цією метою умов користувача і вбудованої мови STATISTICA, завдяки чому можна сформувати дуже гнучкі умови відбору спостережень до вибірки.

Застосовуючи алгоритми, наведені у статті, дослідники, навіть добре не обізнані з роботою пакету, можуть достатньо просто сформувати необхідні вибіркові сукупності для вирішення практичних завдань.

Ключові слова: вибірковий метод, вибірка, вибіркова сукупність, підмножина, статистичний пакет STATISTICA, STATISTICA.

Вступ. Швидке зростання потужностей сховищ даних дозволило накопичувати величезні обсяги даних, які зараз прийнято називати терміном Big Data. Аналіз великих даних дозволяє виявити ринкові тенденції, уподобання клієнтів та багато іншого. Недарма наразі спостерігається справжнє полювання за цими даними, а провідні світові корпорації намагаються за будь-яку ціну одержати доступ до них.

Зрозуміло, що для роботи з такими даними потрібно мати як обізнаних фахівців-аналітиків, так і відповідний інструментарій. Потужним інструментом аналізу даних є пакет статистичного аналізу даних STATISTICA, навички роботи з яким повинні мати всі фахівці, що спеціалізуються на аналізі великих даних.

Досить часто користувачам-аналітикам потрібні не всі дані, а певна їх частина, наприклад якогось регіону, вікової групи і т. ін. Тому вони повинні бути обізнаними з можливостями STATISTICA для відбору потрібних даних.

STATISTICA має низку спеціальних засобів, що дозволяють відібрати дані, до яких належать такі інструменти як фільтрація даних, створення вибірок і формування підмножин.

Найбільш гнучким засобом відбору даних, що надає користувачеві можливості для відбору даних, є формування підмножин. Разом із тим він вимагає від користувача певних знань, оскільки за цим варіантом здійснюється формування вибірки із застосуванням умов користувача і вбудованої мови STATISTICA.

Метою даної роботи було проаналізувати засоби статистичного пакету STATISTICA для формування вибіркових сукупностей з існуючих робочих таблиць.

Для досягнення мети необхідно було розв'язати такі завдання:

1. Розглянути механізм створення вибіркових сукупностей за допомогою підмножин.

2. Розглянути застосування умов користувача і вбудованої мови STATISTICA для створення гнучких варіантів відбору даних для підмножин.

3. Розглянути механізм створення вибіркових сукупностей за допомогою механізму створення випадкової вибірки за різними варіантами.

4. Надати прості і зрозумілі приклади формування вибіркових сукупностей.

Огляд літератури. Серед українських статистиків, що розглядають питання вибіркового методу, широко відомі О. Гладун, А. Єріна, З. Пальян, М. Пугачова. Низка українських статистиків розглядають роботу з пакетом STATISTICA. Серед них можна назвати знову А. Єріну, а також Д. Єріна, В. Данілова.

Разом із тим, автор досі ніде не зустрічав розгляд практичного формування вибіркових сукупностей за допомогою підмножин з використанням STATISTICA. Дана робота допоможе статистикам-аналітикам з'ясувати алгоритми формування на конкретних прикладах.

Методи дослідження. У статті розглядаються та аналізуються алгоритми формування вибіркових сукупностей засобами пакету STATISTICA двома основними способами: шляхом формування підмножин (в тому числі із застосуванням умов користувача і вбудованої мови STATISTICA) і за допомогою механізму створення випадкової вибірки різними варіантами з існуючих робочих таблиць даних.

Результати. У статті на зрозумілих прикладах надано опис алгоритмів формування підмножин з використанням вбудованої мови STATISTICA і умов користувача. Розглянуто приклад формування такої підмножини. Розглянуто алгоритм формування вибірок, у тому числі досить складного характеру. Надані поради відносно аспектів формування вибірок.

Застосовуючи алгоритми, наведені у статті, дослідники, навіть добре не обізнані з роботою пакету, можуть достатньо просто сформувати необхідні вибіркові сукупності для вирішення своїх завдань.

Розглянемо процес формування підмножини. Застосуємо для цього наступний умовний приклад. Є дані про доходи від допоміжної діяльності готелю, що надає своїм постояльцям низку послуг, у тому числі послуги салону краси. Аналітика цікавлять дані про постояльців за 2018 р., яким надавалися послуги у салоні краси, вартість яких становить не менше 300 грн. Дані про допоміжну діяльність готелю знаходяться у таблиці даних, що міститиме три змінні (ознаки), які розташовані у такому порядку: «Вид послуги», «Вартість послуги», «Дата».

Формування підмножини здійснюється за таким алгоритмом.

Створення підмножини починають з вибору або змінної групування або команди створення підмножини. Для створення підмножини виконується команда **Data ▶ Subset**, після чого відкриється вікно «Create a Subset».

Вибір змінної групування може бути здійснений різними способами. За найпростішим варіантом це можна зробити так само, як це робиться в електронних таблицях, тобто натиснувши на імені змінної, яка міститься у першому рядку таблиці з даними. Вибір змінної можна здійснити і пізніше у вікні параметрів модуля, використовуючи кнопку **«Variables»**. Її натискання ініціює появу вікна вибору змінної «Select the variables for the analysis». Після вибору змінної праворуч від кнопки **«Variables»** відображається ім'я вибраної змінної. Оскільки для створення підмножини потрібні всі три змінні, то і вибрати потрібно всі три.

На початку формування підмножини користувач має визначитися, чи планує він надалі працювати зі сформованою вибіркою чи вона потрібна йому тільки одноразово. За першим варіантом у вікні створення підмножини «Create a Subset» слід залишити прапорець біля поля-мітки «Create new spreadsheet», що приведе до створення нової таблиці з існуючих первинних даних.

Для формування умов відбору слід натиснути кнопку **«Cases»**, що призведе до появи вікна «Spreadsheet Case Selection Conditions», у якому і задаються умови відбору до вибірки. Щоб задати умови відбору встановлюємо прапорець біля поля-мітки «Enable Selection Conditions», після чого стають доступними поля для завдання умов. В групі полів «Include cases» встановлюємо прапорець біля поля-мітки «Specific, selected by», вказуючи тим самим що відбір буде здійснюватися за певними правилами. Це у свою чергу зробить доступним поле «Expression», що розташоване під полем «Specific, selected by».

Використовуючи спеціальний механізм системи, в тестовому полі «Expression» вводиться умова. Згідно з правилами роботи цього механізму умова формується наступним чином:

1. Ліва частина виразу – це ім'я змінної, що позначається літерою «*v*» із додаванням її порядкового номера у таблиці даних.

2. Після імені змінної вводиться логічний оператор.

3. Права частина умови є виразом, з яким порівнюється змінна в умові. Вираз може містити математичну операцію, вбудовану функцію системи або просто число.

Умов може бути декілька; вони поєднуються звичайними логічними операторами AND, OR, NOT (ТА, АБО, НІ).

Нагадаємо, що користувача цікавлять дані про послуги, які надавалися постояльцям готелю в 2018 році, вартість яких становить не менше 300 грн.

Перша умова стосується змінної «Вид послуги», що є категоріальною. Кожна з її категорій (послуга) має свій код. Припустимо, послуги у салоні краси мають код «104». Враховуючи, що змінна «Вид послуги» у таблиці даних знаходиться у першому стовпчику, умова матиме такий вигляд:

v1 = 104

Друга умова має встановити, що вартість послуги (друга за порядком змінна у таблиці даних) має бути більше 300 грн, що формулюється так:

v2 >= 300

Нарешті третя умова передбачає, що дані мають бути відібрані тільки за 2018 р. Ця умова може бути реалізована за допомогою вбудованої функції системи і має такий вигляд (змінна "Дата" є третьою за порядком у таблиці даних):

DTYEAR(v3) = 2018

Отже, загальна умова виглядає так:

v1 = 104 and v2>300 and DTYEAR(v3) = 2018

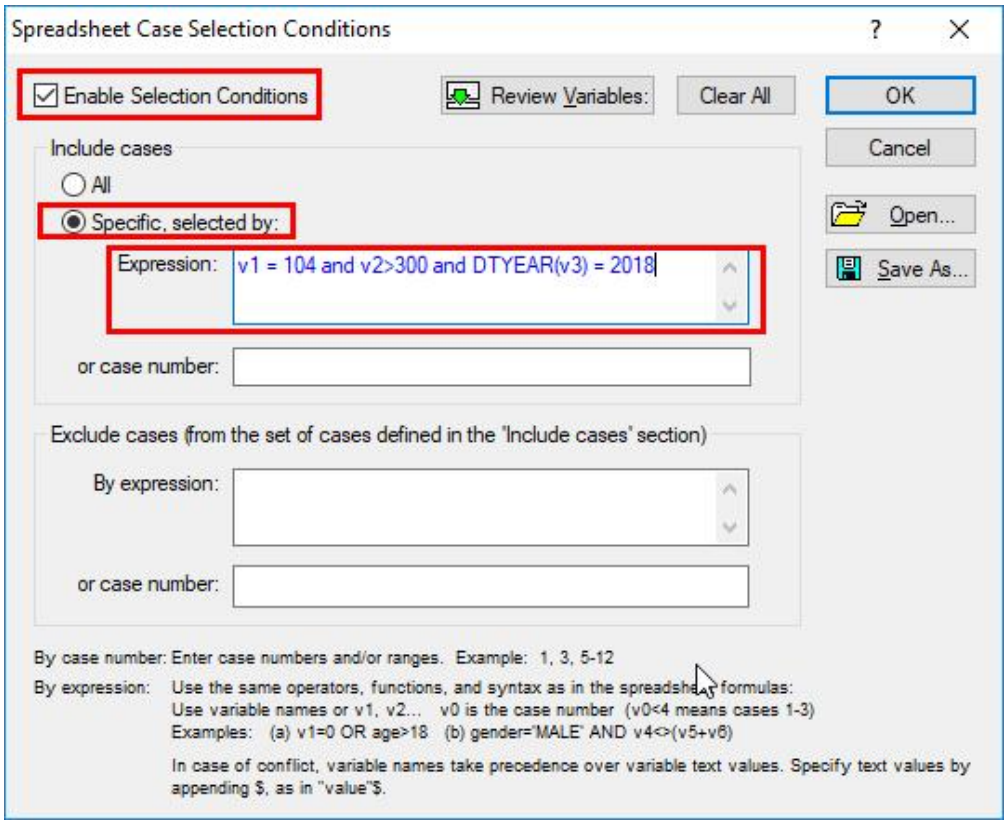


Рисунок 1. ***Приклад формування умов***

Натискаючи кнопку **«ОК»** повертаємось у вікно «Spreadsheet Case Selection Conditions».

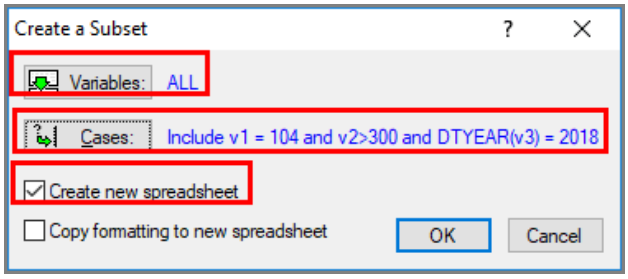


Рисунок 2. ***Відображення заданих параметрів***

У цьому вікні натискаємо кнопку **«ОК»**. Це приведе до створення нової таблиці даних, яка міститиме вибірку (підмножину) , сформовану за критеріями користувача.

Таблиця 1

Фрагмент таблиці даних зі сформованою підмножиною

	Вид послуги	Вартість, грн	Дата
1	Салон краси	310,00 грн	07.03.2018
2	Салон краси	400,00 грн	08.03.2018
3	Салон краси	310,00 грн	08.03.2018

Знання роботи вище описаного механізму дозволить користувачу формувати вибіркові сукупності практично за будь-якими критеріями.

Ще одним способом формування вибірок у програмі є формування вибіркових сукупностей.

STATISTICA надає можливість формування вибірок з існуючих робочих таблиць даних, при цьому це можна зробити різними способами.

Для створення вибірки слід виконати команду **Data ▶ Sampling** (Створити випадкову вибірку), після чого відкриється однойменне вікно «Create a Random Sampling».

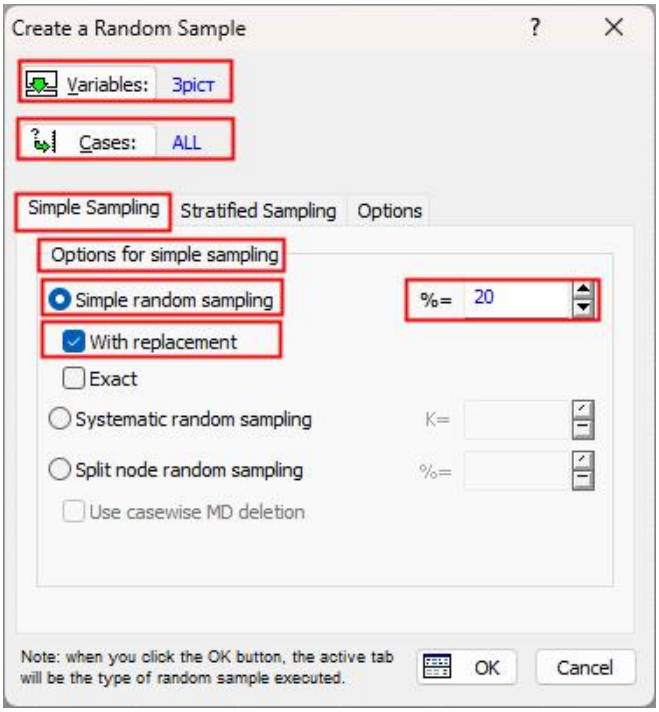


Рисунок 3. ***Створення вибірки. Завдання параметрів***

Процес формування вибірки починається з визначення змінних та (або) спостережень, з яких буде створюватися вибіркова сукупність. За замовчуванням вибірка буде формуватися з усіх спостережень (Cases: ALL), а нова таблиця з даними міститиме всі змінні таблиці, з якої формується вибірка. Якщо у новій таблиці з даними не потрібні певні змінні, слід скористатися кнопкою **«Variables» (Змінні)** і вибрати ті змінні, які потрібно залишити у вибірці.

У цьому ж вікні визначається спосіб формування вибірки і задаються додаткові параметри її формування. Вибір способу визначається на вкладці «Simple Sampling» (Простий вибір) у групі «Options for simple sampling» (Опції для простої вибірки).

За будь-яким способом поле-мітка «With Replacement» (Вибір з поверненням) визначає метод відбору елементів вибірки, який, як відомо, може бути повторним (встановлюється прапорець біля цього поля) або безповторним (прапорець не встановлюється).

Розглянемо способи формування вибіркової сукупності, що надає програма.

Власне випадковий або простий випадковий

У групі «Options for simple sampling» слід встановити перемикач в положення «Simple random sampling» (Проста випадкова вибірка) і в полі «%=» визначити відсоток спостережень з таблиці у вибірці.

Натискання кнопки **«ОК»** спричинить створення нової таблиці з даними, яка буде містити визначений відсоток спостережень поточної робочої таблиці з даними.

Систематичний (механічний)

У групі «Options for simple sampling» слід встановити перемикач в положення «Systematic random sampling» (Систематичний випадковий вибір), а в полі «К»

визначити крок – число, що буде визначати кількість елементів, які необхідно пропускати у таблиці даних (генеральній сукупності) до наступного елемента, що буде потрапляти до вибірки. Перший елемент вибірки відшукується випадковим чином серед перших К спостережень таблиці даних.

Натискання кнопки **«ОК»** спричинить створення нової таблиці з даними.

Розподілений випадковий

Такий варіант формування вибірки відсутній серед прийнятих у статистичній практиці, які були описані вище. За цим варіантом таблиця з даними розділяється на дві таблиці, відсоток яких визначається в полі « %=». Використання даного варіанту задається у групі «Options for simple sampling" встановленням перемикача в положення «Split node random sampling» (Випадкова вибірка з розділеними групами).

Розшарований

Нагадую, що такого роду відбір елементів сукупності передбачає включення до вибірки елементів з *урахуванням структури* вибірки. У STATISICA застосувати цей метод відбору можна у вікні «Create a Random Sampling» на вкладці «Stratified Sampling» (Стратифікована (розшарована) вибірка). ***Стратифікація*** – це розподіл елементів сукупності на групи за певною ознакою.

Якщо для розглянутих перед цим способів формування вибірки вибір змінних був не потрібен, то для цього способу наявність змінної стратифікації є обов'язковою, тобто у вікні вибору необхідно вибрати змінну. Це є зрозумілим, оскільки дійсно не зрозуміло, за якою саме змінною слід виділяти групи, якщо таких змінних дві або більше.

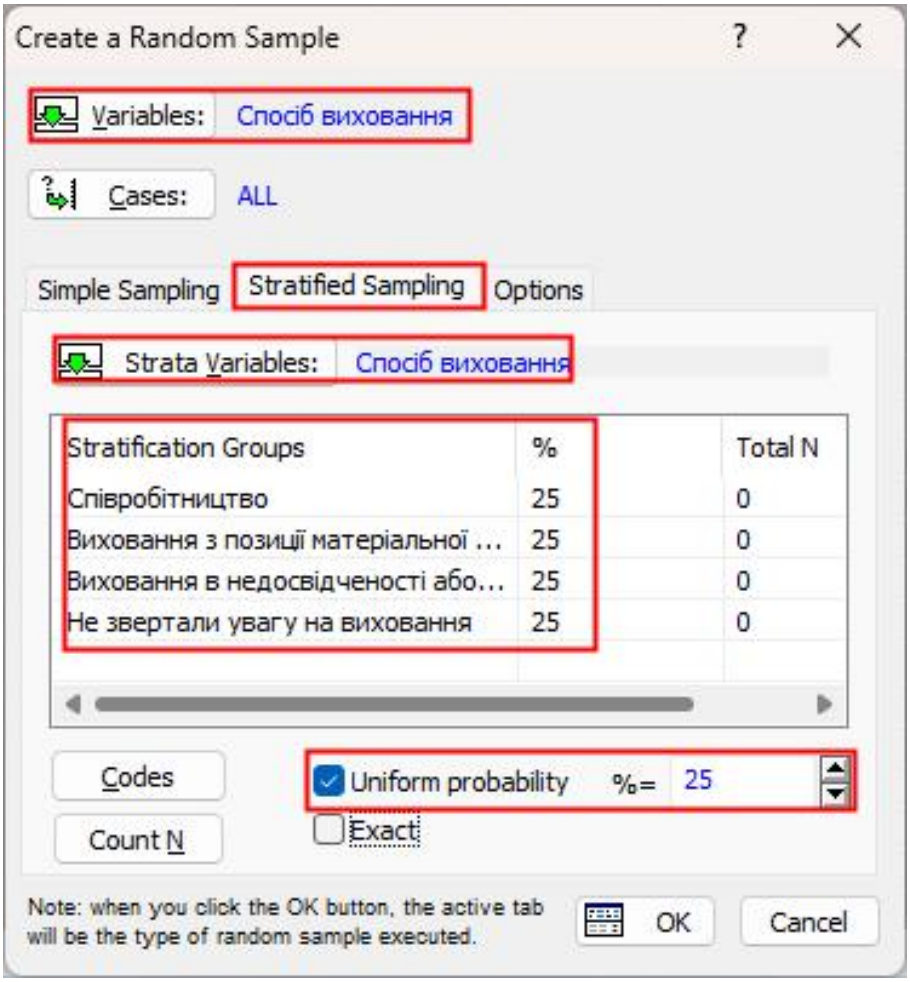


Рисунок 4. ***Створення розшарованої вибірки. Завдання параметрів***

Для її визначення слід натиснути кнопку **«Strata Variables»** (змінні стратифікації) і визначити змінну стратифікації.

Надалі визначаються групи, за якими буде здійснюватися процес стратифікації. Як правило, це всі групи.

Кількість груп, що відбирається до вибірки, визначається двома полями-мітками: «Uniform Probability» (Рівні ймовірності) і «Exact» (Точне число).

Встановлення прапорця поруч з полем-міткою «Uniform Probability» визначає, що з кожної групи до вибіркової сукупності буде відбиратися кількість елементів сукупності *приблизно* пропорційно її долі у робочій таблиці з даними. Для прикладу, наведеного вище, кількість груп становить 40, 60, 60 і 40 спостережень, або 20 %, 30 %, 30 % і 20 %. Ця чисельність визначається автоматично після натискання кнопки **«CountN»**.

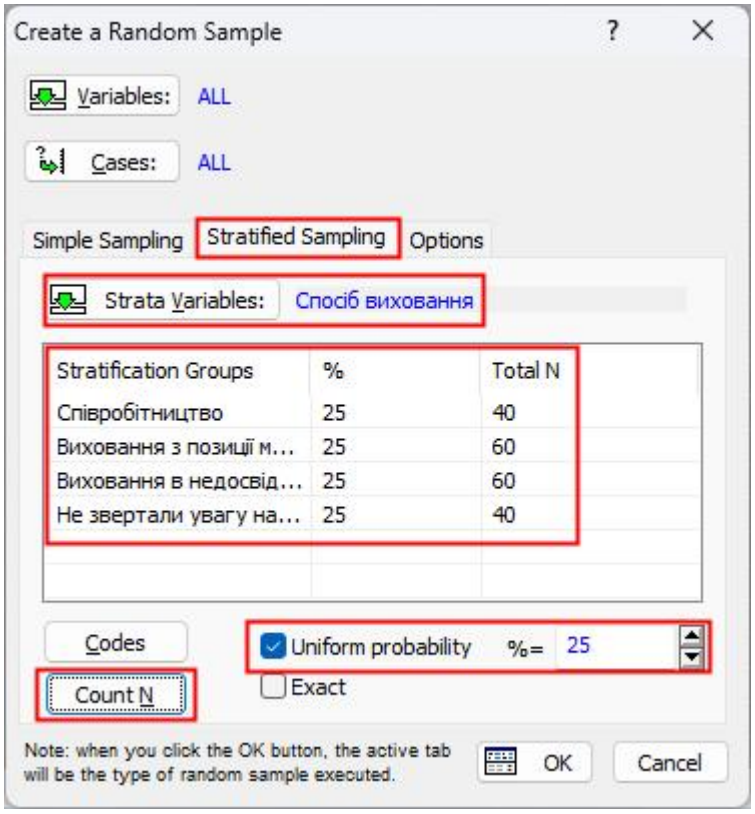
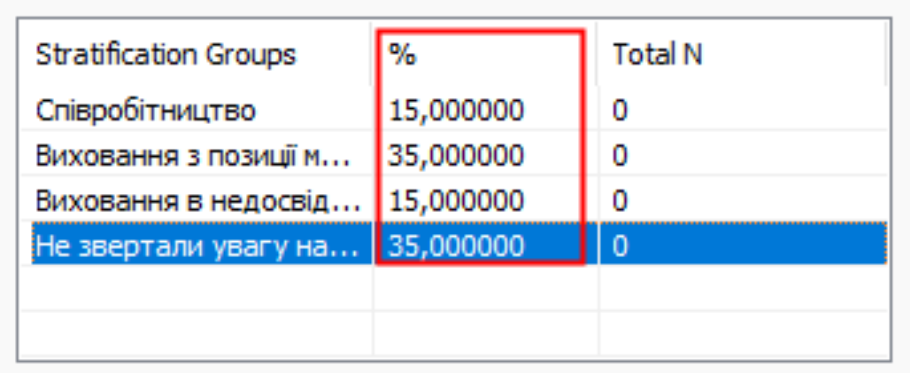


Рисунок 5. ***Результати створення розшарованої вибірки***

Важливо пам'ятати, що встановлення відбору для *всіх* груп в полі «%=» певного значення, наприклад, 25 %, не буде означати, що до вибірки потрапить 10, 15, 15 і 10 спостережень. Для кожної групи ці значення кожного разу будуть іншими, досить відчутно коливаючись, наприклад 9, 18, 16 і 13, а потім 11, 13, 15 і 12 (фактично експериментально одержані автором дані). Більше того, простий підрахунок показує, що навіть чисельність вибірки буде різною: для першої вибірки – 57, а для другої – 51.

За будь-яким варіантом відбору слід визначити відсоток спостережень кожної групи, що буде потрапляти до вибірки. Це число бути загальним, воно задається в полі «%=». Але воно може бути і унікальним для кожної групи. За таким варіантом воно вводиться праворуч від назви групи у стовпчику «=».



Stratification Groups	%	Total N
Співробітництво	15,000000	0
Виховання з позиції м...	35,000000	0
Виховання в недосвід...	15,000000	0
Не звертали увагу на...	35,000000	0

Рисунок 6. ***Створення розшарованої вибірки. Визначення відсотків для груп***

Зауважу, що за таким варіантом *поле* «%=» не заповнюється.

Поле-мітка «Exact» (точно) встановлює, що до кожної групи має потрапити кількість спостережень, яка є чітко пропорційною її долі у робочій таблиці з даними. Для нашого прикладу чисельність груп становить 40, 60, 60 і 40 спостережень. Якщо встановити відбір для всіх груп в полі «%=», наприклад, 25 %, то це не буде означати, що до вибірки потрапить 10, 15, 15 і 10 спостережень відповідно з кожної групи.

З переліку додаткових параметрів, що розташовані на вкладці «Options» вікна «Create a Random Sampling», можна відмітити «Use Diehard-certified random number generator» (Використовуйте генератор випадкових чисел, сертифікований Diehard). Він визначає можливість підключення власного випадкового генератора системи, який, як стверджується, є дуже якісним. Він працює більш повільно, про що, власне кажучи, навіть вказано у назві показника: «note: this algorithm is slower» (зауважу, що цей алгоритм працює повільніше), тому його не рекомендують використовувати для великих масивів даних, але для відносно невеликих за обсягом даних (до 1000 спостережень), автор не відчув жодного уповільнення під час формування вибірки.

Обговорення

У науковій літературі мало висвітлено питання формування вибіркових сукупностей за допомогою наявного у STATISTICA механізму підмножин, коли потрібно створювати достатньо складні умови. У даному дослідженні показано, що для формування таких складних умов потрібно вміти використовувати як умови користувача, так і вбудовану мови STATISTICA.

Висновки. Під час проведеного дослідження на зрозумілих прикладах розглянуто способи формування вибіркових сукупностей, у тому числі досить складного характеру, як шляхом створення підмножин, так і за допомогою механізму створення випадкової вибірки за різними варіантами з існуючих робочих таблиць даних. Розглянуто формування підмножин з використанням умов користувача і вбудованої мови STATISTICA.

Використовуючи наведені у статті приклади формування вибірок, дослідники, навіть добре не обізнані з роботою пакету STATISTICA, можуть достатньо просто сформувати необхідні вибіркові сукупності для вирішення своїх завдань.

Література

1. Фетісов В. С. Пакет статистичного аналізу даних STATISTICA: навчальний посібник. Ніжин: Видавництво НДУ ім. М. Гоголя, 2018. 102 с.

2. StatSoft. Офіційний сайт. URL: https://www.statistica.com/en/software/statistica-evaluation.

References

1. Fetisov, V.S. (2018). Paket statystychnoho analizu danykh STATISTICA [STATISTICA statistical data analysis package]. Nizhyn: Vydavnytstvo NDU im. M.Hoholia [in Ukrainian].

2. StatSoft. Official website. URL: https://www.statistica.com/en/software/statistica-evaluation.

Fetisov V.
PhD in Economics,
Associate Professor of the Department of Information Technologies,
Physics, Mathematics and Economics,
Nizhyn Mykola Gogol State University
fetisov.vs@ndu.edu.ua
orcid.org/0000-0003-4316-9060

PRACTICAL QUESTIONS OF SAMPLES FORMATION IN STATISTICA

The modern world is a world of information. Using the Internet provides access to huge amounts of various data. However, this data becomes truly useful only after the existing patterns in it are identified, which allows it to be used in practical activities.

It is clear that understanding the essence of data can be revealed only after its statistical analysis. Analyzing large amounts of data requires the use of computing technology. But using general-purpose software like Excel does not always solve the problem. In this case, researchers must be able to use specialized software – statistical data processing packages, among which there are two undisputed leaders – STATISTICA and SPSS. This article discusses working with STATISTICA.

It is not always possible to work with the entire data set (and sometimes it is not necessary at all) and one has to use a part of the data, i.e. a sample population, on the basis of which conclusions are drawn regarding the entire (general) population as a whole.

The issues of working with a sample population in statistics are discussed in a separate section – the sampling method. Among the Ukrainian statisticians who consider the issues of the sampling method are O. Gladun, A. Yerina, Z. Palian, and M. Pugacheva. Some of Ukrainian statisticians consider working with the STATISTICA package. Among them we can name again A. Yerina, as well as D. Yerina, V. Danilov. At the same time, the author has not yet encountered a discussion of the sampling method using STATISTICA.

Quite often, a researcher must to select the part of the data from the general population. Methods of forming sample populations are one of the sections of the sampling method. Algorithms for solving this problem are discussed in the article.

***The purpose of the article**. To consider the options for forming sample sets using the STATISTICA package, including its built-in language.*

***Practical significance**. The article describes the algorithms for forming subsets using the built-in STATISTICA language with clear examples. An example of forming such a subset is considered. The algorithm for generating samples, including those of a rather complex nature, is considered. Advice on the aspects of sample formation is given.*

***Conclusions**. Using the algorithms presented in the article, researchers, even those who are not well acquainted with the package, can easily form the necessary sample sets to solve their problems.*

Key words: sampling method, sample, sample population, subset, statistical package STATISTICA, STATISTICA.