УДК 004.45:[311.12:519.25]
JEL Classification: C83, L86
Doi: 10.31767/su. 4(83)2018.04.14

В. С. Фетісов,
кандидат економічних наук, доцент,
доцент кафедри інформаційних технологій та аналізу даних,
Ніжинський державний університет імені Миколи Гоголя,
E-mail: fetisval@gmail.com

Побудова групувань з використанням пакета STATISTICA

Викладено механізм побудови групувань у пакеті статистичного аналізу STATISTICA, подано приклади побудови таблиці частот та настроювання системи для дискретних і неперервних ознак. Висвітлено механізм функціонування пакета STATISTICA під час застосування параметра "приблизна кількість інтервалів". Розглянуто стандартні механізми обмеження кількості груп і визначення користувачем умов, за яких можна гнучко формувати межі інтервалів, у тому числі для нерівних і відкритих інтервалів. Наведено приклади візуалізації результатів процедури групування.

Ключові слова: *групування даних, побудова групувань, інтервал, частотний аналіз, статистичний пакет STATISTICA, таблиця частот.*

Сучасні технології дозволяють накопичувати великі обсяги даних, які нині є об'єктом пильної уваги з боку багатьох структур. Аналіз великих даних дозволяє виявити ринкові тенденції, уподобання клієнтів та багато іншого. Але такі обсяги первинних статистичних даних просто фізично не дозволяють здійснювати їх аналіз вручну. Вкрай важко виконати навіть просте групування кількох тисяч спостережень, здійснити найпростіші розрахунки статистичних показників, не кажучи вже про застосування складних статистичних методів.

Усі ці питання успішно вирішуються за допомогою спеціального класу програмного забезпечення – прикладних статистичних пакетів, які широко застосовуються у практичній діяльності під час роботи з первинними даними у найрізноманітніших галузях. Серед незаперечних лідерів на світовому ринку програмного забезпечення цього класу зазначимо універсальний пакет STATISTICA, що має потужні можливості й дозволяє застосовувати широке коло статистичних методів. Досить часто їх застосуванню передує зведення даних статистичного спостереження, одним з основних елементів якого є їх групування. Незважаючи на відносну простоту цього етапу обробки даних, ручний процес їх зведення може зайняти тривалий час під час опрацювання великих масивів даних, не кажучи вже про дуже високу ймовірність виникнення помилок. Тому застосування для групування даних статистичних пакетів є зрозумілим і доцільним. Зазначимо, що оскільки на ринку представлені як англомовна, так і російськомовна версії системи, то інтерфейс і команди пакета наводяться у статті у двох варіантах.

Сам процес групування даних у пакеті статистичного аналізу STATISTICA виконується зовні достатньо просто. Але під час практичної роботи у користувача можуть виникнути певні проблеми, пов'язані, зокрема, з побудовою інтервалів. На думку автора, ця проблема відома практично всім, хто здійснює групування даних в пакеті, водночас вона зовсім не висвітлюється як у спеціалізованій літературі, так і в Інтернеті. До того ж користувачі достатньо часто не знають про можливості, що надає пакет при побудові таблиць частот.

Побудову групування (як, власне кажучи, і реалізацію практично кожного методу в пакеті) варто починати з вибору або змінної групування, або статистичного методу, що в термінології пакета STATISTICA називається аналізом. Побудова таблиці частот є складовою модуля "Bases Statistics/Tables" ("Основные статистики и таблицы"). Сам модуль, як, до речі, і будь-який інший, можна завантажити кількома способами:

1. Виконати команду Statistics ▶ Bases Statistics/Tables (Анализ ▶ Основные статистики и таблицы).
2. На панелі "Statistics" ("Анализ") натиснути кнопку  "Bases Statistics/Tables" ("Основные статистики и таблицы").
3. Натиснути кнопку  "Start menu" ("Вызвать меню часто используемых средств") і вибрати з меню Statistics ▶ Bases Statistics/Tables (Анализ ▶ Основные статистики и таблицы).

За будь-яким варіантом виконання з'явиться вікно "Basic Statistics and Tables" ("Основные статистики и таблицы"). Для побудови групування у вікні вибирається пункт "Frequency tables" ("Таблицы частот"), після чого з'явиться однойменне вікно. Воно містить кілька вкладок з різноманітними параметрами-настройками, основною серед яких є "Advanced" ("Дополнительно").

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 назву змінної, що міститься у першому рядку таблиці з даними. Проте можна і не виділяти повністю змінну, достатньо просто встановити курсор у клітинці з потрібною змінною або виділити клітинку у рядку. Якщо буде виділено кілька змінних, то для кожної з них у робочій книзі (результати аналізу) буде створено окремий аркуш з групуванням. Вибір змінної можна здійснити і пізніше у вікні параметрів модуля, використовуючи кнопку "Variables" ("Переменные"). Її натискання ініціює появу вікна вибору змінної "Select the variables for the analysis" ("Выберите переменные для анализа"). Після вибору змінної праворуч від кнопки "Variables" відображається ім'я вибраної змінної.

Після вибору змінної для побудови таблиці частот натискаємо кнопку "Summary" ("ОК"), яка для зручності відображається на кожній вкладці вікна таблиці частот, що й ініціює побудову останньої. При цьому система розраховує частоти для кожного значення змінної. Крім цього, за замовчуванням розраховуються також частки, кумулятивні частоти і кумулятивні частки. Але настройки системи дозволяють створювати інтервали для змінної за різними алгоритмами і значно розширювати результатну інформацію, про що і піде мова у статті.

Алгоритм побудови групувань розглядається на умовному прикладі даних про доходи від допоміжної діяльності готелю "Зірка", що надає своїм постояльцям такі послуги, як збереження речей у камерах схову, послуги хімчистки та салону краси, масаж тощо. Таблиця даних міститиме дві змінні (ознаки), одна з яких є дискретною ("Вид послуги"), інша – неперервною ("Вартість послуги").

Побудова групування для дискретної ознаки здійснюється максимально просто. Вибрав змінну, ініціюємо розрахунок натисканням кнопки "Summary" ("ОК"). Групування здійснюється за параметрами, що містяться у вікні модуля на вкладці "Advanced" ("Дополнительно") у групі "Categorization method for tables & graphs" ("Метод категоризации для таблиц и графиков"). За замовчуванням створюється окрема група для кожного значення змінної, що визначається встановленням значення перемикача вибору метода групування в положення "All distinct value" ("Все различные значения").

Під час розрахунку система автоматично встановлює п'ять знаків після коми для значень у графі "Percent" (Процент) і чотири знаки – у графі "Comulative – Percent" (Кумулятивный процент). Зрозуміло, що пакет STATISTICA не має загальної настройки для кількості знаків після коми, оскільки для кожного показника може знадобитися власна точність. Відсутня така настройка і серед налаштувань аналізу. Але користувач все ж таки має можливість встановити потрібну кількість знаків після коми. Для цього у таблиці частот слід виділити стовпчики "Percent" і "Comulative – Percent". У межах таблиці викликається контекстне меню і виконується команда Format (Формат) ▶ Cells (Ячейки). Це спричиніть появу вікна "Format Cells" (Формат ячеек), у якому в полі "Decimal places" (Число десятичных знаков) і задається потрібне число. На жаль, такий варіант не дозволяє надалі зафіксувати для всіх інших таблиць частот задану кількість знаків після коми. Разом із тим користувач може запам'ятати будь-який власний варіант формату даних, вибравши з контекстного меню команду Format (Формат) ▶ New From Selection (Новый формат). Надалі цей формат запам'ятовується у списку активних форматів команди Format (Формат) і його можна швидко застосувати не тільки для будь-якої таблиці частот, а й узагалі для будь-яких таблиць з результатами.

Перший стовпчик таблиці містить значення групувальної ознаки (категорія у термінології програми), табл. 1. Якщо це значення є кодом, якому поставлено у відповідність текстовий еквівалент (текстова мітка у термінології програми), то під час виведення можна замінити значення коду на текст. Для цього у вікні параметрів модуля на вкладці "Advanced" ("Дополнительно") слід встановити прапорець для полямітки "with text label" ("с текстовыми значениями"), унаслідок чого таблиця частот набуває такий вигляд (табл. 2).

Таблиця 1

Таблиця частот для змінної "Вид послуги". Категорія – код

Category	Count	Cumulative – Count	Percent	Cumulative – Percent
101,0000	14	14	32,6	32,6
102,0000	8	22	18,6	51,2
103,0000	5	27	11,6	62,8
104,0000	14	41	32,6	95,3
Missing	2	43	4,7	100,0

Таблиця 2

Таблиця частот для змінної "Вид послуги". Категорія – текстова мітка

Вид послуги	Count	Cumulative – Count	Percent	Cumulative – Percent
Камера схову	14	14	32,6	32,6
Хімчистка	8	22	18,6	51,2
Масаж	5	27	11,6	62,8
Салон краси	14	41	32,6	95,3
Missing	2	43	4,7	100,0

За замовчуванням біля поля-мітки “with text label“ (“с текстовыми значениями”) у вікні параметрів міститься прапорець, а отже, значення змінної за наявності текстових міток будуть відображатися саме у текстовому вигляді. Останній рядок таблиці даних “Missing” (Пропущені) містить пропущені (відсутні) дані.

Механізм використання системою пропущених даних потрібно знати (на що користувачі не завжди звертають увагу) і вміти використовувати. Настройка механізму роботи з пропущеними даними регулюється перемикачем “MD deletion (Missing data)” (“Удаление ПД (пропущенных данных)”), розташованим у правому нижньому куті вікна настройок модуля. Завдяки ньому STATISTICA уможливлює користувачеві різні варіанти роботи з відсутніми даними. Як правило, такі дані просто відкидаються з подальших розрахунків. Але користувач може встановити режим, за яким здійснюється імпутація даних, наприклад середніми значеннями змінної у модулі множинної регресії.

Під час побудови таблиці частот користувач має можливість встановити один з двох режимів роботи з відсутніми даними:

1. Casewise (Построчное). Вилучаються спостереження з відсутніми даними для будь-якої з вибраних змінних.

2. Pairwise (Парами). Вилучаються спостереження з відсутнім значенням тільки для змінної групування. Для інших змінних пропуски дозволяються.

Розглянемо дію настройки механізму роботи з пропущеними даними для нашого прикладу, де таблиця даних містить три спостереження з відсутніми даними (табл. 3).

Таблиця 3

Спостереження з відсутніми даними

Вид послуги	Вартість, грн	Дата
Камера схову		07.03.2018 р.
	50,50	07.03.2018 р.
		07.03.2018 р.

Якщо групування будується тільки для однієї змінної, то за будь-яким значенням перемикача “MD deletion (Missing data)” (“Удаление ПД (пропущенных данных)”) у таблиці результатів рядок з пропущеними даними (останній рядок таблиці) завжди відображатиме два відсутні значення. Отже положення перемикача не впливає на результати під час побудови групування для однієї змінної. Його використання стає доцільним під час побудови групування для двох і більше змінних одночасно. У цьому випадку, як зазначалося раніше, для кожної змінної у робочій книзі створюється окремий аркуш з групуванням. Якщо під час побудови кількох таблиць частот перемикач “MD deletion (Missing data)” (“Удаление ПД (пропущенных данных)”) має значення “Casewise” (“Построчное”), то будуть вилучені усі спостереження, де є відсутні дані, незалежно від того, для якої змінної вони відсутні, що демонструє табл. 4. Якщо ж перемикач має значення “Pairwise” (“Парами”), то для кожного групування буде відкидатися по два спостереження, де відсутні дані саме для змінної групування.

Таблиця 4

Відсутні дані у робочій книзі з двома таблицями частот

Вид послуги	Count	Cumulative – Count	Percent	Cumulative – Percent
Камера схову	13	13	30,2	30,2
Хімчистка	8	21	18,6	48,8
Масаж	5	26	11,6	60,4
Салон краси	14	40	32,6	93,0
Missing	3	43	7,0	100,0

У модулі за замовчуванням розраховуються групові та кумулятивні частоти і проценти. Додаткові показники, що можна розраховувати, визначаються на вкладці “Options” (“Опции”).

STATISTICA дозволяє швидко побудувати гістограму групування натисканням кнопки “Histograms” (“Гистограммы”), що розташована на перших двох вкладках вікна параметрів, рис. 1

Побудовану гістограму можна редагувати, змінюючи багато її параметрів, доступ до яких здійснюється викликом контекстного меню в межах гістограми. Редагувати можна також її окремі елементи. Наприклад, виділивши назву осі, користувач одержує доступ до параметрів осі, де він може, зокрема, змінити її назву.

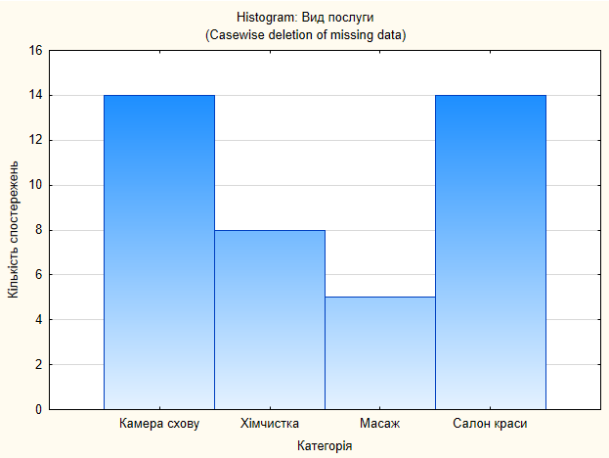

Рис. 1. Гістограма розподілу за категоріями послуг

Користувач має можливість виключати з групування окремі групи, застосовуючи стандартний механізм відбору груп, реалізований у пакеті. Скористатися ним можна на вкладці "Advanced" ("Дополнительно"), натиснувши кнопку "Specific Grouping Codes" ("Заданные группирующие коды"), після чого з'являється вікно вибору кодів "Select codes for the selected variables". Вікно має дві кнопки – "All" ("Все") і "Zoom" ("Инфо"). Зрозуміло, що натискання першої кнопки приводить до вибору всіх груп. Натискання кнопки "Zoom" ("Инфо") ініціює появу вікна з усіма групами, в якому звичайним чином здійснюється вибір потрібних груп. Таблиця частот після цього міститиме тільки відібрані групи, яких, до речі, повинно бути не менше двох.

Механізм побудови групування для неперервної ознаки розглянемо на тих самих даних, використовуючи з цією метою зміну "Вартість послуги". Як було згадано вище, групування даних здійснюється за параметрами, розташованими у вікні модуля на вкладці "Advanced" ("Дополнительно") у групі "Categorization method for tables & graphs" ("Метод категоризации для таблиц и графиков"). При цьому за замовчуванням створюється окрема група для кожного значення змінної, що визначається встановленням перемикача вибору метода групування в положення "All distinct value" ("Все различные значения "). Фрагмент таблиці, одержаної при цьому, наведено нижче (табл. 5).

Таблиця 5

Таблиця частот для змінної "Вартість послуги"

Category	Count	Cumulative – Count	Percent	Cumulative – Percent
50,00 грн	1	1	2,3	2,3
50,50 грн	5	6	11,6	14,0
100,00 грн	1	7	2,3	16,3
...				
400,00 грн	1	40	2,3	93,0
450,00 грн	1	41	2,3	95,3
Missing	2	43	4,7	100,0

STATISTICA має низку параметрів побудови групування, знання та використання яких дозволяє одержати зручніші результати або взагалі відібрати певну частину даних. Наприклад, вибір метода групування "Integer Categories" ("Целые интервалы (категории)") ініціює побудову групування тільки для цілих значень змінної із відкиданням решти спостережень (табл. 6).

Таблиця 6

Таблиця частот для змінної "Вартість послуги" для цілих значень

Category	Count	Cumulative – Count	Percent	Cumulative – Percent
50,00 грн	1	1	2,7	2,7
100,00 грн	1	2	2,7	5,4
...				
400,00 грн	1	34	2,7	91,9
450,00 грн	1	35	2,7	94,6
Missing	2	37	5,4	100,0

Для неперервних ознак групування будується, як правило, за інтервалами. Під час побудови такого групування пакет STATISTICA надає можливість використовувати різні варіанти визначення інтервалів. Це здійснюється на вкладці "Advanced" ("Дополнительно") у групі "Categorization method for tables & graphs" ("Метод категоризации для таблиц и графиков"). Параметри цієї групи надають користувачу такі можливості побудови інтервалів:

1. Задати певну кількість рівних інтервалів у полі "No. of exact intervals" ("Число равных интервалов").

2. Побудувати групування з приблизною" кількістю інтервалів, для чого використовується параметр "Neat" intervals; approximate no." ("Приблизительное число интервалов"). За таким варіантом початкове значення нижньої межі першого інтервалу визначається, виходячи з мінімального та максимального значен-

ня змінної. Останньою цифрою меж інтервалів і кроку інтервалу є цифри 0, 1, 2 або 5. При цьому слід мати на увазі, що фактична кількість інтервалів може не збігатися із заданою, на що, власне кажучи, вказує сама назва параметра: приблизна (approximate) кількість.

3. Визначити крок інтервалу, встановивши перемикач у положення "Step size" ("Размер шага") і задати крок інтервалу. Початкове значення нижньої межі першого інтервалу автоматично визначається системою як мінімальне з усіх значень, але й може бути задано користувачем у полі "starting with" ("начать с"). Для застосування першого варіанта достатньо встановити прапорець для поля-мітки "at minimum" ("с минимального значения"), а для другого – ввести потрібне значення. Якщо початкове значення нижньої межі першого інтервалу більше за окремі значення змінної, то спостереження з такими значеннями просто відкидаються.

Розглянемо варіанти побудови групування з п'яті рівних інтервалів змінної "Вартість послуги". Зазначимо, що мінімальне значення змінної становіть 50 грн, максимальне – 450 грн. Результати ручного зведення подано у табл. 7.

Таблиця 7

Розподіл вартості послуг від додаткової діяльності готелю "Зірка"

Вартість, грн	Кількість послуг
50–130	17
130–210	11
210–290	9
290–370	2
370 і більше	2
Разом	41

За першим варіантом задаємо п'ять рівних інтервалів у полі "No. of exact intervals" ("Число равных интервалов") і одержуємо такі результати (табл. 8).

Таблиця 8

Таблиця частот для змінної "Вартість послуги" для п'яті рівних інтервалів

From To	Count	Cumulative – Count	Percent	Cumulative – Percent
0,00 грн <x<=100,00 грн	7	7	16,3	16,3
100,00 грн <x<=200,00 грн	19	26	44,2	60,5
200,00 грн <x<=300,00 грн	11	37	25,6	86,0
300,00 грн <x<=400,00 грн	3	40	7,0	93,0
400,00 грн <x<=500,00 грн	1	41	2,3	95,3
Missing	2	43	4,7	100,0

Пакет STATISTICA у цьому випадку (хоча це і не є правилом) призначає нижній межі першого інтервалу значення "0". Сам же варіант зовсім не відповідає ряду розподілу з табл. 1. Отже, такий варіант побудови інтервалів доцільний, коли потрібно починати формування меж інтервалів саме з нульового значення.

Аналогічний результат матимемо при використанні параметра "Neat" intervals; approximate no." ("Приблизительное число интервалов"). Зрозуміло, що результати для обох параметрів однакові, оскільки в обох випадках визначені системою межі інтервалів будуть кратні 10, тобто "круглому" числу.

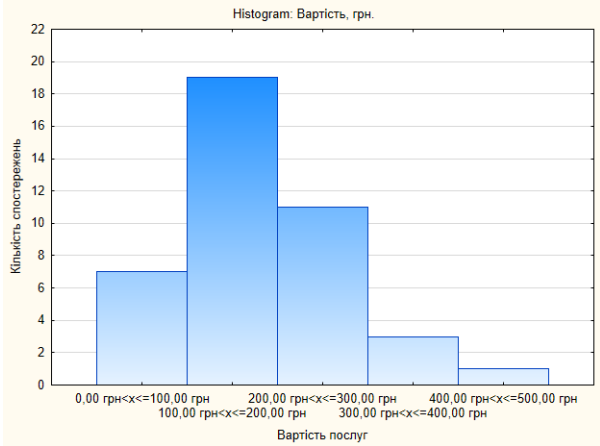

Рис. 2. Гістограма розподілу за вартістю послуг

Як і для дискретної ознаки, за обома варіантами пакет STATISTICA дозволяє побудувати гістограму групування натисканням кнопки "Histograms" ("Гистограммы"), що розташована на перших двох вкладках вікна параметрів (рис. 2).

При використанні параметра "Neat" intervals; approximate no." ("Приблизительное число интервалов") система будує межі інтервали і вибирає крок так, щоб остання цифра для значень меж інтервалів була закруглена до простих значень з останньою значущою цифрою 1, 2 або 5. Це має забезпечити кращу читабельність даних групування. Відправною точкою для визначення меж і кроку інтервалів є мінімальне та максимальне значення змінної. Орієнтуючись на задану кількість інтервалів, система підбирає крок, кратний 1, 2 або 5 так, щоб фактична кількість не перевищувала заданої. Зрозуміло, що фактична кількість інтервалів при цьому, як зазначалося раніше, може не збігатись із заданою.

За визначеним у такий спосіб значенням кроку система визначає кількість інтервалів, яких, до речі, не може бути менше трьох. Зауважимо, що досить часто фактична кількість інтервалів для досить широкого діапазону значень параметра “Neat” intervals; approximate no.” (“Приблизительное число интервалов”) буде однаковою для тих самих мінімального і максимального значення змінної. Так, для нашого прикладу фактична кількість інтервалів буде однаковою за варіації значення параметра від 3 до 9 і має становити п’ять, оскільки максимальне значення (450) для групування з кроком 100 потрапляє до п’ятого інтервалу. При цьому система (і це дещо незрозуміло) завжди створює на один інтервал більше, тобто у нашому випадку система сформує шість інтервалів.

Кращому розумінню формування меж інтервалів при використанні параметрів “No. of exact intervals” (“Число равных интервалов”) і “Neat” intervals; approximate no.” (“Приблизительное число интервалов”) сприяє приклад, за яким групування повинно містити сім груп за рівних (табл. 9) і приблизних (табл. 10) інтервалів. Цей приклад ілюструє відмінності зазначених варіантів: рівних інтервалів будується саме 7, приблизних – 6.

Таблиця 9

Таблиця частот для змінної “Вартість послуги” для семи рівних інтервалів

From To	Count	Cumulative – Count	Percent	Cumulative – Percent
17 грн <x<=83 грн	6	6	14,0	14,0
83 грн <x<=150 грн	13	19	30,2	44,2
150 грн <x<=216,6667	9	28	20,9	65,1
217 грн <x<=283 грн	9	37	20,9	86,0
283 грн <x<=350 грн	2	39	4,7	90,7
350 грн <x<=417 грн	1	40	2,3	93,0
417 грн <x<=483 грн	1	41	2,3	95,3
Missing	2	43	4,7	100,0

Таблиця 10

Таблиця частот для змінної “Вартість послуги” для семи приблизних інтервалів

From To	Count	Cumulative – Count	Percent	Cumulative – Percent
0,00 грн <x<=100,00 грн	7	7	16,3	16,3
100,00 грн <x<=200,00 грн	19	26	44,2	60,5
200,00 грн <x<=300,00 грн	11	37	25,6	86,0
300,00 грн <x<=400,00 грн	3	40	7,0	93,0
400,00 грн <x<=500,00 грн	1	41	2,3	95,3
500,00 грн <x<=600,00 грн	0	41	0,0	95,3
Missing	2	43	4,7	100,0

Прийнятим у статистиці варіантом групування, що відповідає результатам, наведеним у табл. 1, є застосування параметра “Step size” (“Размер шага”).

Отже, потрібно побудувати групування з п’яті рівних інтервалів. Оскільки крок інтервалу дорівнює 80, встановлюємо перемикач в положення “Step size” (“Размер шага”) і задаємо в полі праворуч від нього значення 80. Оскільки мінімальне (початкове значення першого інтервалу) становить 50, то для визначення нижньої границі першого інтервалу можна або ввести це мінімальне значення в полі “starting with” (“начать с”), або перекласти завдання визначення мінімального значення на систему, встановивши прапорець для поля-мітки “at minimum” (“с минимального значения”).

Натиснувши кнопку “Summary” (“ОК”), одержуємо таблицю частот (табл. 11).

Таблиця 11

Таблиця частот для змінної "Вартість послуги" при застосуванні параметра "Step size"

From To	Count	Cumulative – Count	Percent	Cumulative – Percent
50,00 грн <=*x*<130,00 грн	17	17	39,5	39,5
130,00 грн <=*x*<210,00 грн	11	28	25,6	65,1
210,00 грн <=*x*<290,00 грн	9	37	20,9	86,0
290,00 грн <=*x*<370,00 грн	2	39	4,7	90,7
370,00 грн <=*x*<450,00 грн	1	40	2,3	93,0
450,00 грн <=*x*<530,00 грн	1	41	2,3	95,3
Missing	2	43	4,7	100,0

Як видно з табл. 11, система будує на один інтервал більше, ніж під час ручної побудови, тобто створює шостий інтервал 450–530 для значення 450. І так вона буде робити завжди. Суто математично це є правильним, оскільки значення 450 знаходиться на межі двох інтервалів і має бути віднесено до нижньої межі наступного інтервалу. За ручним варіантом групування ця проблема вирішувалась дуже просто: останній інтервал створюється відкритим: "370 і більше".

STATISTICA надає можливість користувачеві вирішити проблему "зайвого" інтервалу, але розв'язується вона вручну. Для цього на вкладці "Advanced" ("Дополнительно") слід встановити перемикач вибору методу групування в положення "User-specified categories" ("Определенные пользователем категории") і натиснути кнопку  , що знаходиться праворуч від цього перемикача. Це спричинить появу вікна "Define Categories" ("Определить категории"). А надалі користувач самостійно формує кількість інтервалів і їх межі, використовуючи спеціальний механізм системи, за якого кожний інтервал створюється за умовою, що, своєю чергою, складається з двох частин. Перша частина – це заздалегідь визначені у системі правила, наприклад "Include If" ("Включать, если"), "Include Cases" ("Верные наблюдения"). Друга частина є виразом для правила, що формується так.

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

2. Після імені змінної вводиться логічний оператор.

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

Отже, залишаємо для всіх інтервалів (категорій) правило "Include If" ("Включать, если"). Формуємо умови для кожної категорії. Наприклад, перший інтервал 50–130 описуємо як *v*2<130 (у таблиці даних змінна "Вартість послуги" є другою за порядком розташування), другий інтервал 130–210 – як *v*2>=130&*v*2<210, а останній інтервал 370 і більше – як *v*2>=370. Такий варіант дозволяє створювати закриті, відкриті, нерівні інтервали, усуваючи всі проблеми під час побудови групування.

За результатами настроювання одержуємо групування (табл. 12) у вигляді, що повністю збігається з даними табл. 7, одержаними під час ручного зведення.

Таблиця 12

Таблиця частот для змінної "Вартість послуги" при застосуванні спеціального механізму

Category	Count	Cumulative – Count	Percent	Cumulative – Percent
Include *v*2<130	17	17	39,5	39,5
Include *v*2>=130&*v*2<210	11	28	25,6	65,1
Include *v*2>=210&*v*2<290	9	37	20,9	86,0
Include *v*2>=290&*v*2<370	2	39	4,7	90,7
Include *v*2>=370	2	41	4,7	95,3
Not selected	2	43	4,7	100,0

Звичайно, такий варіант значно більш трудомісткий порівняно з автоматичною побудовою таблиці частот, але він дозволяє гнучко будувати групування. Разом із тим, якщо потрібно буде ще неодноразово будувати групування із саме такими інтервалами, то можна просто запам'ятати створені умови. У вікні "Define Categories" ("Определить категории") натискаємо кнопку "Save" ("Сохранить") і зберігаємо умови у вигляді файлу з розширенням INI. У подальшому для застосування збережених умов у вікні достатньо просто натиснути кнопку "Open" ("Открыть") і вибрати збережений файл.

Отже застосування механізму визначення користувачем категорій вирішує кілька важливих проблеми:

1. Створення відкритих інтервалів.
2. Формування нерівних інтервалів, що також неможливо здійснити за стандартними настройками.
3. Збереження визначених користувачем меж інтервалів з можливістю подальшого їх використання.

Наведені у статті відомості допоможуть користувачам більш ефективно застосовувати процес групування (побудову частот) пакету STATISTICA. Звичайно, формат статті не надає можливості висвітлити усі можливості цього процесу. Так, представляє інтерес перевірка даних на нормальність як у цілому для даних, так і для окремих груп, побудова діаграми розмаху та інші операції.

Список використаних джерел

1. Фетісов В. С. Пакет статистичного аналізу даних STATISTICA. Ніжин: Вид-во НДУ імені Миколи Гоголя, 2018. 102 с.
2. Statsoft. Электронный учебник по статистике. URL: http://statsoft.ru/home/textbook/default.htm

References

1. Fetisov, V. S. (2018). *Paket statystychnoho analizu danykh STATISTICA. [Package of statistical data analysis STATISTICA].* Nizhyn: Vydavnytstvo NDU imeni Mykoly Hoholia [in Ukrainian].
2. Statsoft. Elektronnyi uchebnik po statistike. [Electronic textbook on statistics]. *statsoft.ru.* Retrieved from http://statsoft.ru/home/textbook/default.htm [in Russian].

В. С. Фетисов,
кандидат экономических наук, доцент,
доцент кафедры информационных технологий и анализа данных,
Нежинский государственный университет имени Николая Гоголя

Построение групп с использованием пакета STATISTICA

Изложен механизм построения групп в пакете статистического анализа STATISTICA, представлены примеры построения таблицы частот и настройки системы для дискретных и непрерывных признаков. Освещены механизм функционирования пакета STATISTICA при применении параметра "приблизительное количество интервалов". Рассмотрены стандартные механизмы ограничения количества групп и определения пользователем условий, при которых можно гибко формировать границы интервалов, в том числе для неравных и открытых интервалов. Приведены примеры визуализации результатов процедуры группирования.

Ключевые слова: *группировка данных, построение групп, интервал, частотный анализ, статистический пакет STATISTICA, таблица частот.*

V. S. Fetisov,
PhD in Economics, Associate Professor,
Associate Professor of Department for Information Technology and Data Analysis,
Nizhyn Mykola Gogol State University

Constructing Groupings by Use of STATISTICA Software Package

STATISTICA software package for statistical analysis incorporates a wide range of advanced statistical methods. Quite often they are preceded by aggregating statistical survey data, which main component is their grouping. Although this phase of statistical data processing is relatively simple, the manual process of aggregation can be time-consuming given the need to process large data arrays, not mentioning a high probability of errors. Therefore, the all-purpose STATISTICA software package is a logical and reasonable tool for grouping of data.

The article shows the grouping algorithm in STATISTICA software package, with focus on setup when constructing tables of frequencies of discrete and continual characters. Various options of grouping are scrutinized, with providing examples of their visualization.

A large number of STATISTICA parameters offers ample opportunities for constructing user tables, but users often are not aware of these options or do not know how they can be applied. Yet, the apparently simple grouping process in STATISTICA software package can sometimes require the knowledge of fine mechanisms for its setup. The article gives a detailed description of the mechanisms for creating interval margins when applying the parameter "approximate number of intervals".

The standard algorithm for selection is analyzed, allowing a user to limit the number of groups in a grouping. STATISTICA allows for using a number of grouping parameters, enabling to produce more convenient results or filter them. Thus, setting the clicker for label field "Grouping" in the position "Integer Categories" (integer intervals (categories)) initiates the grouping only for integer values of a variable, by excluding the observations containing its fractional values.

When only standard parameters are used, it will be impossible to form uneven or open intervals. This issue is out of focus in specialized literature and Internet sources. The article shows the algorithm for constructing open intervals by user-set conditions and the process of creating these conditions. This option allows for forming both closed and open intervals by solving all the problems in time of grouping. Because creating such conditions is time consuming, they should be preserved if they are required for further use.

Setting up of STATISTICA software with missing data is analyzed. Its application will be advisable when a grouping for two or more variables is constructed. In this case, a separate sheet with a grouping is to be created in the worksheet for each variable.

Key words: *data grouping, grouping construction, frequency analysis, STATISTICA software package, table of frequencies.*

Бібліографічний опис для цитування:

Фетісов В. С. Побудова групувань з використанням пакета STATISTICA // Статистика України. 2018. № 4. С. 121–129.