СТАТИСТИКА

УДК 519.23

ВИКОРИСТАННЯ PSPP ПІД ЧАС СТАТИСТИЧНОГО АНАЛІЗУ

USE OF THE PSPP IN STATISTICAL ANALYSIS

Остапенко Я.О.
кандидат економічних наук,
доцент кафедри статистики та математичних методів в економіці,
Університет державної фіскальної служби України

У статті висвітлено основні методи статистичного аналізу: описової статистики, кореляційного аналізу, методи регресійного аналізу. Продемонстровано застосування статистичних методів із використанням програми PSPP. Описано основні характеристики та можливості програми PSPP. Доведено ефективність використання програми PSPP для статистичного аналізу.

Ключові слова: *статистичний аналіз, методи статистичного аналізу, описова статистика, кореляційно-регресійний аналіз, статистичні програмні продукти, програма PSPP.*

В статье отражены основные методы статистического анализа: описательной статистики, корреляционного анализа, методы регрессионного анализа. Продемонстрировано применение статистических методов с использованием программы PSPP. Описаны основные характеристики и возможности программы PSPP. Доказана эффективность использования программы PSPP для статистического анализа.

Ключевые слова: *статистический анализ, методы статистического анализа, описательная статистика, корреляционно-регрессионный анализ, статистические программные продукты, программа PSPP.*

The article covers the main methods of statistical analysis: descriptive statistics, correlation analysis, methods of regression analysis. The application of statistical methods using the PSPP program is shown. The main features and features of the PSPP program are described. The efficiency of using the PSPP program for statistical analysis is proved.

Key words: *statistical analysis, methods of statistical analysis, descriptive statistics, correlation-regression analysis, statistical software products, PSPP program.*

Постановка проблеми у загальному вигляді та її зв'язок із важливиминауковими чи практичними завданнями. Прийняття управлінських рішень та будь-яке наукове дослідження базуються на достовірній, репрезентативній, несуперечливій інформації про суб'єкт підприємницької діяльності. Отримати таку інформацію дає змогу економіко-статистичний аналіз із використанням статистичних методів та програмних продуктів. Сучасний ринок програмних продуктів пропонує різноманітні пакети програм для статистичної обробки економічних показників. Питання полягає у виборі ефективних методів аналізу та програмних продуктів, які при цьому не потребували б значних витрат та були простими у використанні.

Аналіз останніх досліджень і публікацій, в яких започатковано розв'язання даної проблеми і на які спирається автор. Використанню програмних продуктів у статистичному аналізі присвячено праці вітчизняних та закордонних науковців: М.В. Роїка, О.І. Присяжнюка, В.О. Денисюка [1], С.А. Айвазян, В.С. Степанова [2], Ж.В. Василенко [3], Р.Є. Майбороди, О.В. Сугакової [4]. Але в наукових працях здебільшого розглядаються статистичні програмні продукти, які потребують значних коштів на придбання та обслуговування, тому питання використання ефективних, але безкоштовних статистичних програм є актуальним.

Виділення невирішених раніше частин загальної проблеми, котримприсвячується означена стаття. Одним з основних етапів наукового дослідження та аналізу діяльності будь-якого суб'єкта підприємницької діяльності є статистичний аналіз. З появою та вдосконаленням сучасних комп'ютерних програм обробки інформації статистична обробка даних піднялася на новий щабель. Тепер аналітик чи дослідник може не мати математичної підготовки. Достатньо оперувати статистичними поняттями та правильно вибрати методи

аналізу. Але в умовах обмежених фінансових ресурсів постає питання ефективного статистичного аналізу за мінімальних витрат, тому на часі є використання безкоштовних статистичних програм, які мають основні статистичні методи обробки показників та прості у використанні. Однією з таких програм є програмне забезпечення PSPP.

Формулювання цілей статті **(постановка завдання)**. Метою статті є розгляд основних етапів та статистичних методів аналізу з використанням безкоштовної статистичної програми PSPP.

Виклад основного матеріалу дослідження з повним обґрунтуваннямотриманих наукових результатів. Розвиток інформаційних технологій та постійне вдосконалення пакетів прикладних програм дають змогу здійснювати пошук необхідної економічної інформації, створювати бази даних, проводити їх швидку та ефективну обробку, здійснювати глибокий аналіз та надавати результати у найбільш зручному вигляді.

Комп'ютерну реалізацію основних статистичних методів обробки даних передбачають електронні таблиці Microsoft Excel, який входить до складу пакету прикладних програм Microsoft Office, QuattroPro та ін.

Більші можливості статистичної обробки мають спеціалізовані пакети STATGRAPHCS, SPSS, SAS, STATISTICA тощо.

Безкоштовним аналогом універсального статистичного пакету SPSS є програмне забезпечення PSPP.

Програмою передбачено повний спектр можливостей статистичного аналізу: частотний аналіз та описова статистика, статистичний аналіз на основі таблиць спряженості, порівняння середніх за t-критерієм Стьюдента та однофакторний дисперсійний аналіз, регресійний аналіз, непараметричні методи аналізу, факторний та кластерний аналізи, аналіз основних компонент тощо.

PSPP призначений для виконання розрахунків так швидко, як це можливо, незалежно від розміру вхідних даних.

Результати статистичних розрахунків та графіки можна зберігати у форматах ASCII, PDF, PostScript, SVG та HTML. Також програмою передбачено широкий діапазон побудови гістограм та різного виду діаграм (стовпчикових, точкових, кругових).

PSPP імпортує дані у форматі електронних таблиць (Gnumeric та OpenDocument) та баз даних PostgreSQL, а також значень, відокремлених комами (CSV), та ASCII. Програма PSPP може імпортувати дані у форматах portable та system SPSS. Доступ до бібліотек, які використовуються PSPP, можна здійснювати на програмному рівні. Інтерфейсом Perl до бібліотек PSPP є PSPP-Perl.

Для роботи в PSPP необхідно створити електронну таблицю або відкрити її, якщо вона створена раніше (рис. 1).

Важливо правильно ввести вхідні дані. У стовпець «Назва» вводиться назва змінної довжиною не більш як вісім символів, якими можуть бути букви і цифри, спеціальні символи. Назва повинна починатися з букви і не може закінчуватися крапкою або «_». Кожна назва унікальна (варто уникати повторів). Назви змінних нечутливі до регістру: писані і друковані, великі та маленькі букви не розрізняються. Не можна використовувати пробіли, знаки інших алфавітів і символи. У стовпчику «Тип» описується тип змінної. Ширина вказує формат стовпця. У стовпці «Знаків після коми» вписують кількість знаків після коми (якщо змінна числового типу). У стовпці «Мітка» вводиться запитання анкети, у стовпці «Мітки

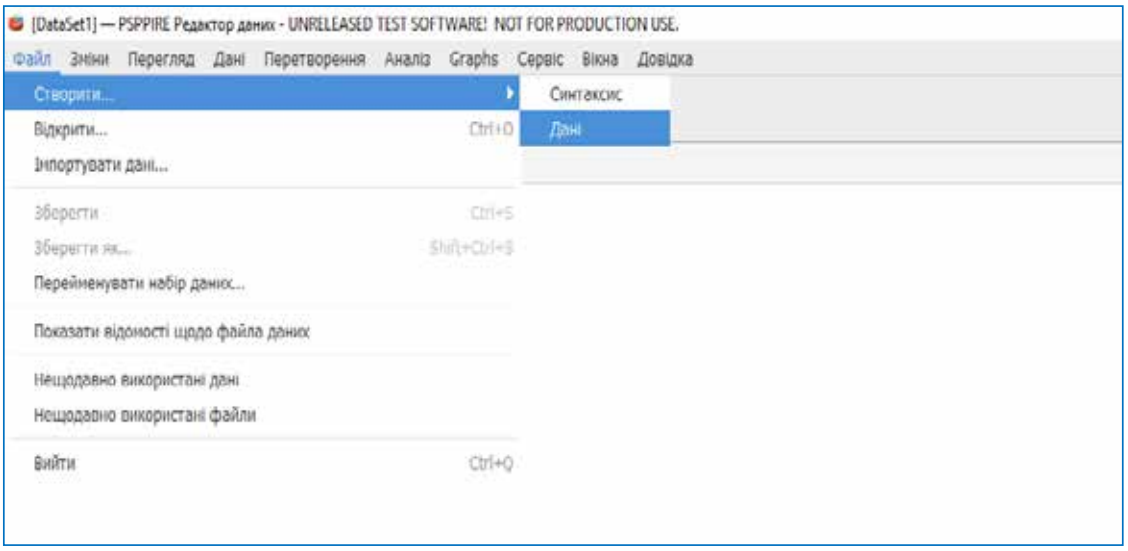

Рис. 1. Створення електронної таблиці в PSPP

Джерело: створено автором

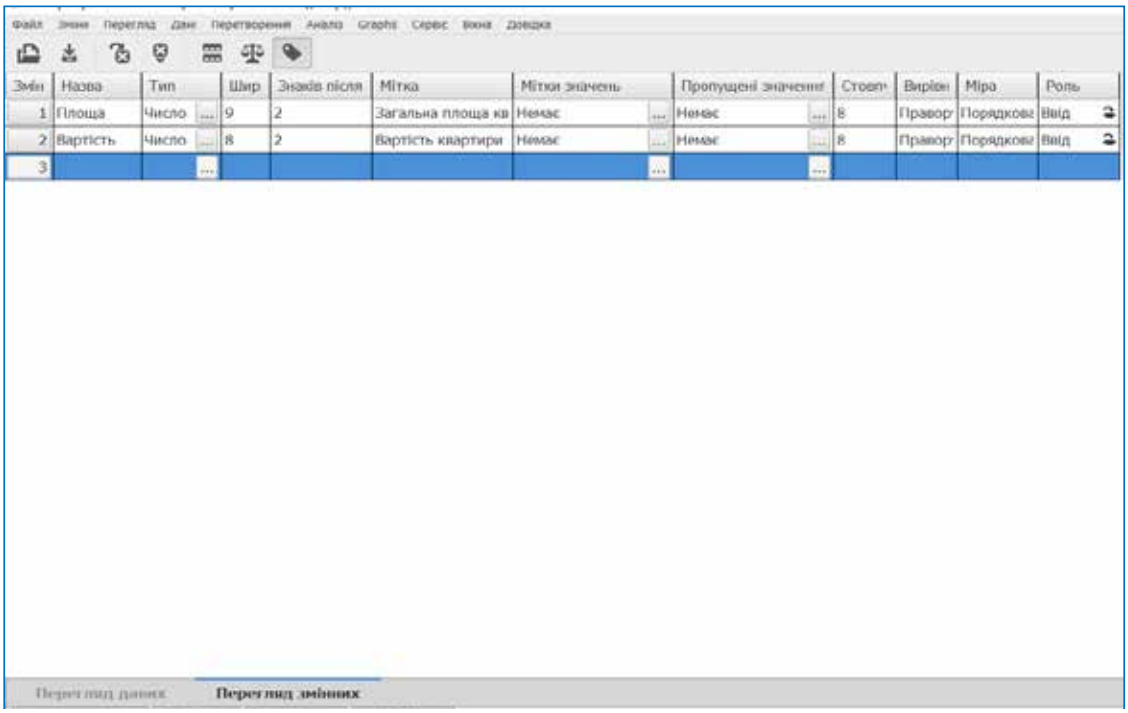

Рис. 2. Введення характеристики змінних

Джерело: створено автором

значень» описується кодування альтернатив відповіді, у стовпці «Пропущені значення» описуються «пусті» клітинки. У стовпці «Стовпчик» задається ширина стовпця. У стовпці «Вирівняти» вирівнюються дані, У стовпці «Міра» вказується рівень вимірювання змінної. Всі ці характеристики змінних описуються у вкладці «Перегляд змінних» (рис. 2).

Заповнення даних відбувається у вкладці «Перегляд даних» (рис. 3).

Першим етапом аналізу даних, що утворюють статистичну сукупність, є розрахунок показників описової статистики, що дають змогу оцінити у цілому, як веде себе досліджувана величина або величини за досить великих обсягів сукупностей об'єктів-носіїв цих величин, тобто проаналізувати характер розподілу цих величин у природі.

Програма PSPP дає змогу розрахувати практично всі застосовувані сьогодні описові статистики, проте слід пам'ятати про те, що для певного типу змінних мають сенс певні показники. Наприклад, відмінності між якісними і кількісними змінними. Різниця між ними – у шкалі виміру. Для якісних даних (дані – конкретні значення змінних) застосовується номінальна шкала, для кількісних – метрична. Показники, які застосовуються для зазначених типів даних, у корені різні. Статистики кількісних змінних незастосовні для якісних, також як і статистики якісних змінних не мають сенсу для кількісних.

Описовими статистиками якісних змінних служать їх частотні характеристики. Вони пока-

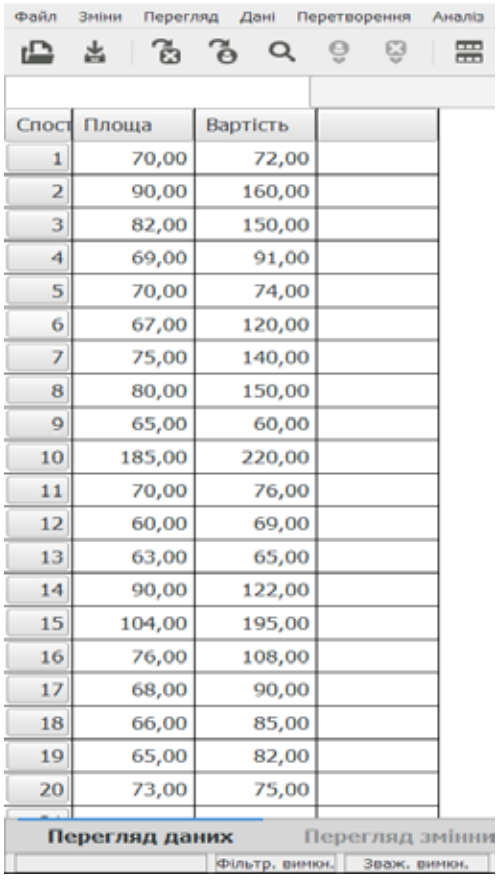

Рис. 3. Уведення даних в електронну таблицю PSPP

Джерело: створено автором

зують структуру розглянутої сукупності, яка розбивається на однорідні за певним значенням показника групи. Для таких груп розглядаються такі показники, як абсолютна чисельність і частка (відсоток, проміле і т. д.) групи в загальному масиві даних.

Створення дискретного ряду розподілу та визначення частот у програмі PSPP здійснюють за алгоритмом «Аналіз → описова статистика → частоти» (рис. 4).

У віконечку «Частоти» необхідно зазначити ознаку для групування (визначення частот) та за потреби можна позначити міткою необхідні розрахунки описової статистики та побудови діаграми (рис. 5).

Результати розрахунків система відобразить окремим файлом, про що повідомить мерехтіння значка PSPP на панелі (рис. 6).

Характер розподілу досліджуваної величини має велике значення для аналізу даних, оскільки той чи інший тип розподілу передбачає застосування тих чи інших статистичних інструментів і показників.

Розрахунок показників описової статистики в PSPP здійснюють за складником вкладки «Аналіз» – «Описова статистика» (рис. 7).

У віконечку «Описова статистика» необхідно зазначити показники для розрахунку (рис. 8).

Результати система відобразить у файлі розрахунків (рис. 9).

У суспільному житті та економічній діяльності часто доводиться вирішувати завдання, спрямовані на виявлення сили і характеру зв'язку між деякими величинами. Величини, що при цьому впливають на інші, називаються факторними ознаками, а величини, на які впливають, – результативними ознаками. Факторні ознаки можуть бути незалежними від дій і рішень компанії (наприклад, стихійні лиха, політичні та економічні рішення керівництва країни, масштабні соціальні процеси і т. д.) або знаходиться в компетенції останньої (цінові, товарні, збутові та інші маркетингові рішення). У першому випадку рішення подібного роду задач може допомогти спрогнозувати поведінку ознаки-результату за відповідного значення факторної ознаки, у другому – вибрати найкращий варіант рішення за факторною ознакою, який міг би дати бажане значення того чи іншого економічного показника (обсягу продажів, споживчих переваг і т. д.).

Метод вирішення подібного роду завдань визначається характером факторної та резуль-

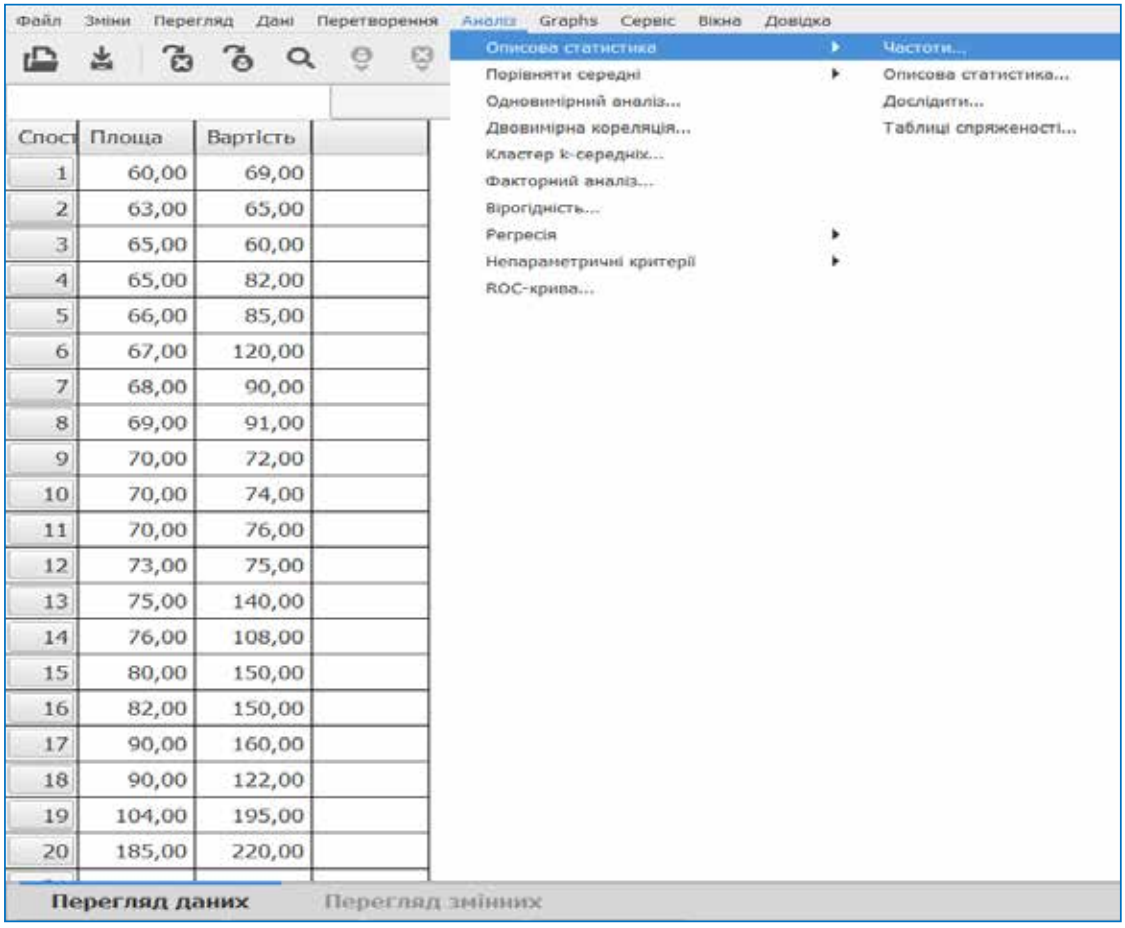

Рис. 4. Виклик віконечка для визначення частот

Джерело: створено автором

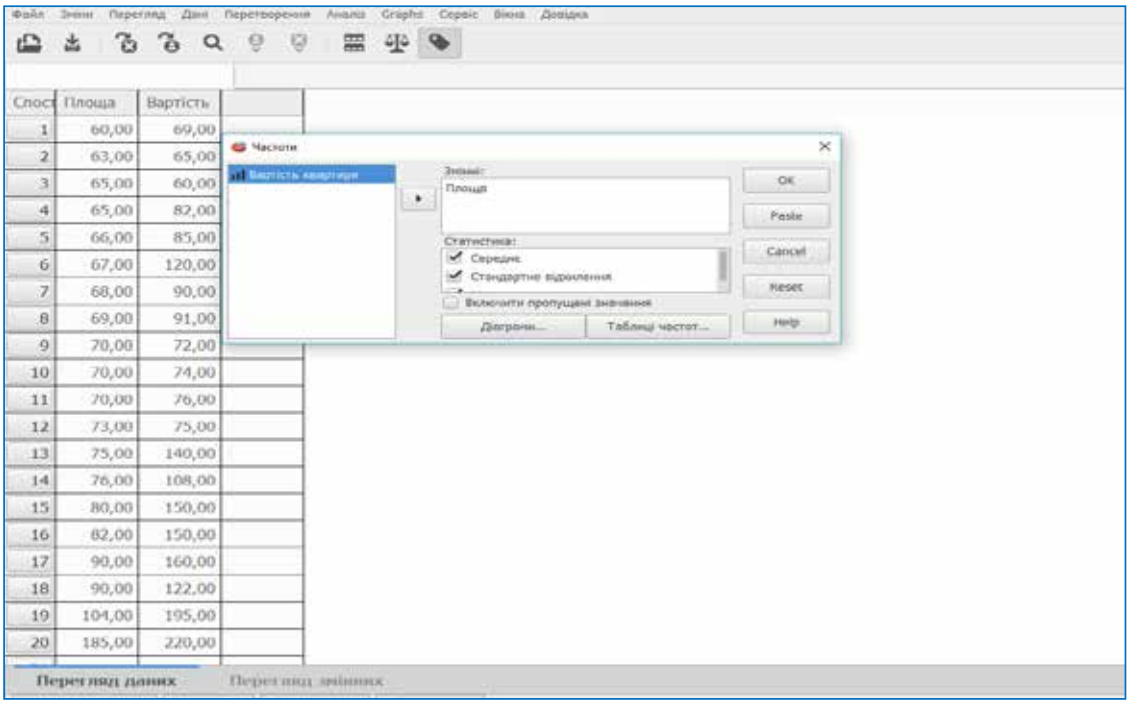

Рис. 5. Побудова дискретного ряду розподілу та визначення показників описової статистики в PSPP

Джерело: створено автором

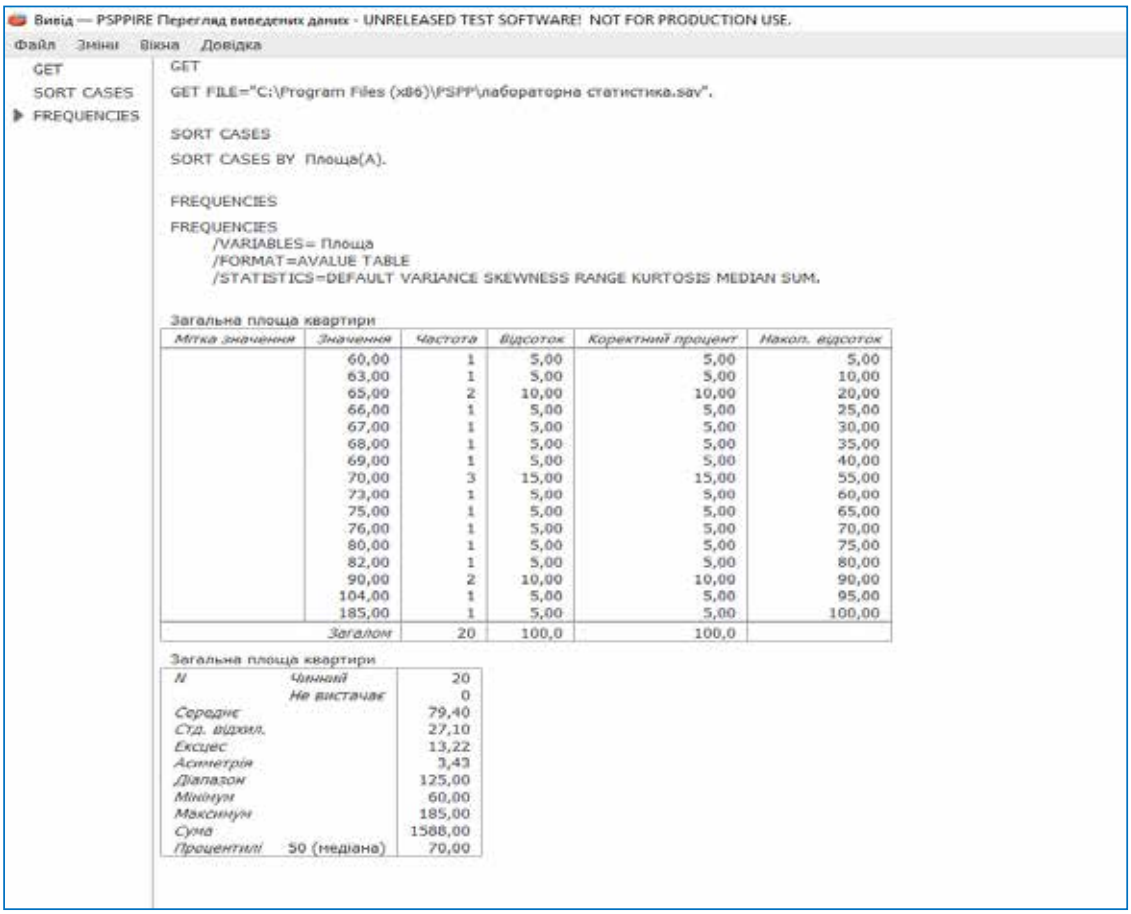

Рис. 6. Результати створення дискретного ряду розподілу в PSPP

Джерело: створено автором

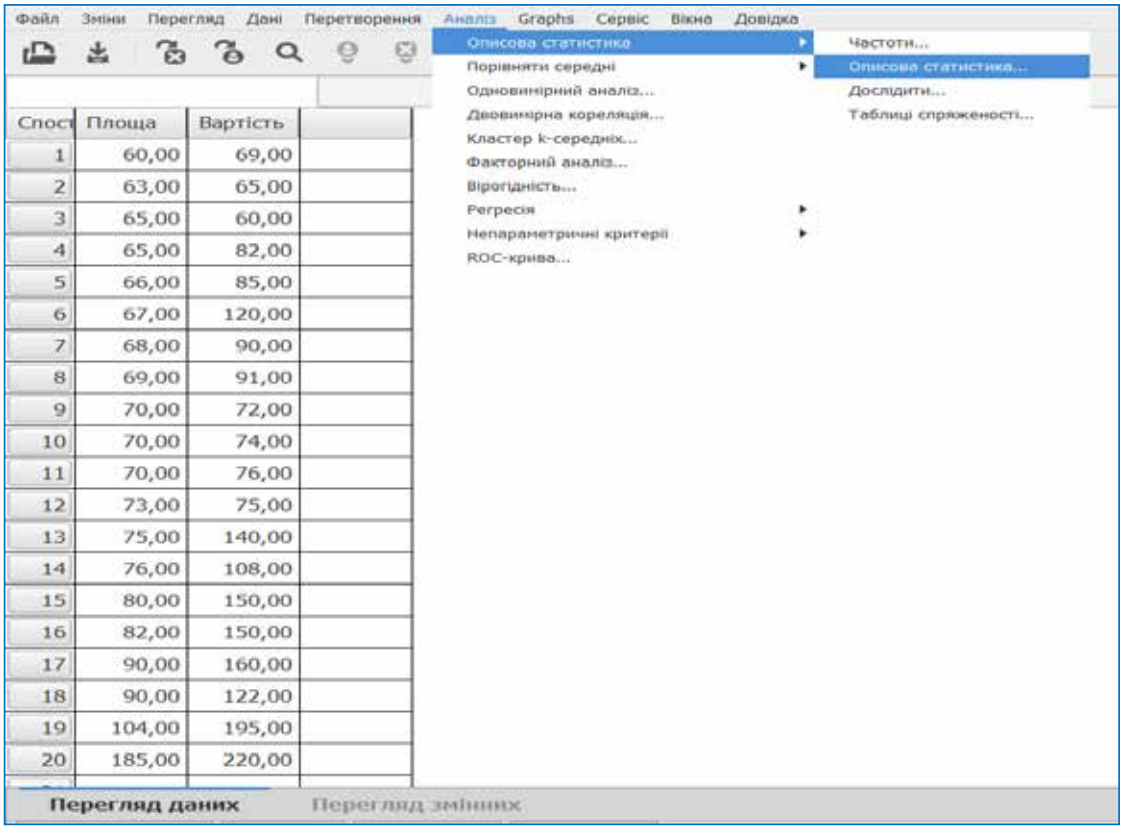

Рис. 7. Використання вкладки «Аналіз» – «Описова статистика» в PSPP

Джерело: створено автором

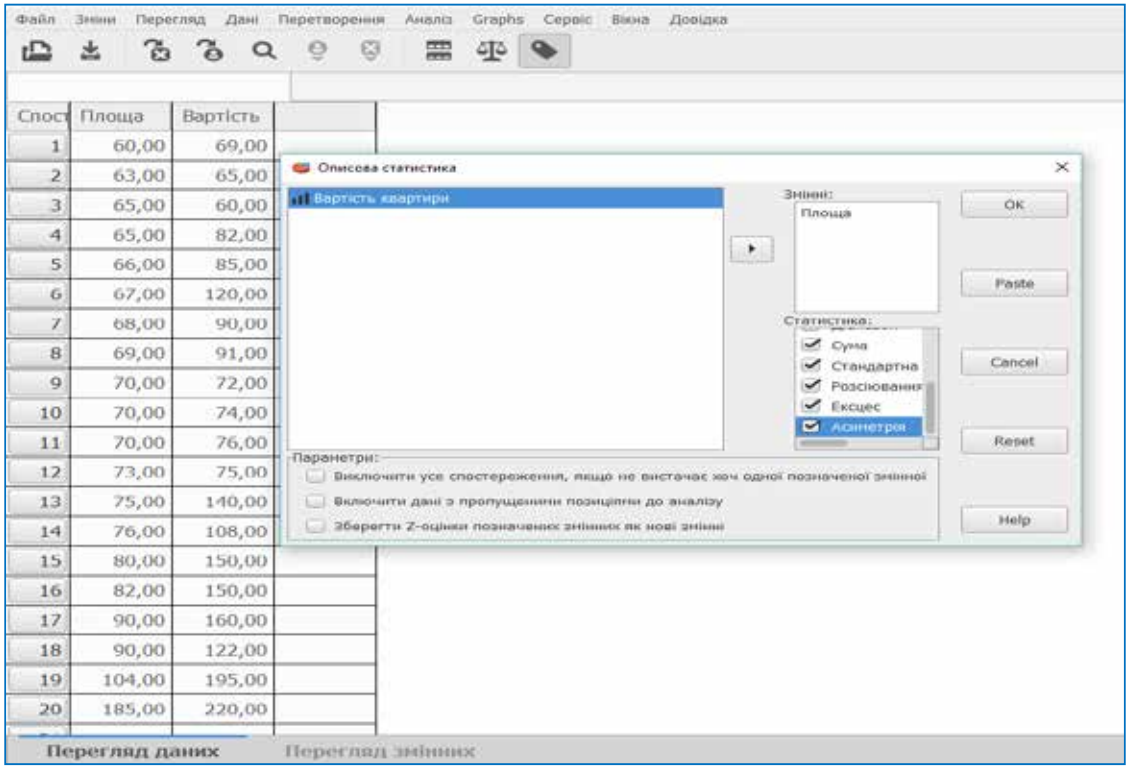

Рис. 8. Заповнення віконечка «Описова статистика» в PSPP

Джерело: створено автором

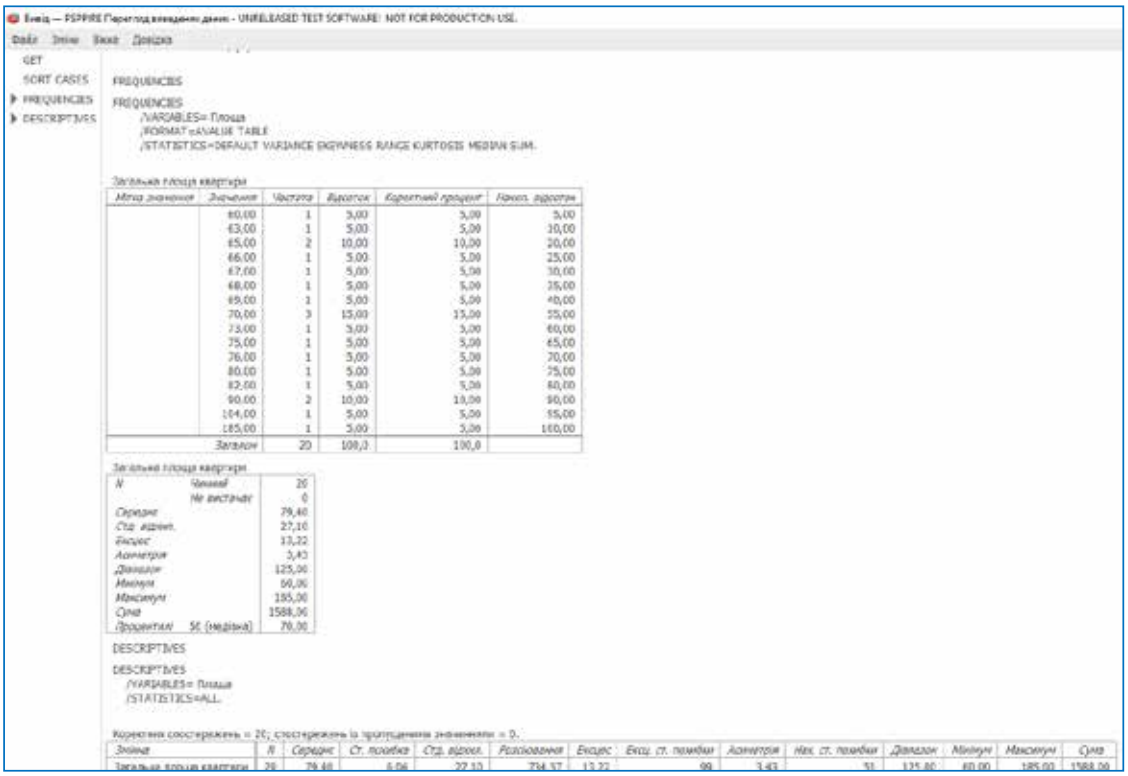

Рис. 9. Показники описової статистики в PSPP

Джерело: створено автором

тативної величин, а саме тим, яку природу вони мають – якісну або кількісну.

Якщо змінні є кількісними показниками, то для аналізу їх взаємозв'язку можна застосувати кореляційно-регресійний аналіз.

Кореляційно-регресійний аналіз, як випливає з назви, складається з двох незалежних етапів – кореляційного та регресійного. Мета першого – виявлення сили взаємозв'язку результативної та факторної змінної (або змінних), другого – виду і параметрів такої залежності.

Силу взаємозв'язку зазвичай оцінюють за допомогою різних показників тісноти зв'язку, серед яких можна виділити непараметричні, або емпіричні: коефіцієнт кореляції рангів (коефіцієнт Спірмена), коефіцієнт Кендала, ранговий коефіцієнт згоди (коефіцієнт конкордації) і коефіцієнт взаємної спряженості Пірсона і параметричні, виведені строго математично: коефіцієнт кореляції знаків (коефіцієнт Фіхнера), коефіцієнт коваріації, лінійний коефіцієнт кореляції Пірсона, коефіцієнт детермінації та емпіричне кореляційне відношення.

Для визначення парного коефіцієнта кореляції в PSPP на вкладці «Аналіз» відзначаємо курсором «Двомірна кореляція» (рис. 10):

Заповнюємо віконечко для обрахування коефіцієнта кореляції (рис. 11).

Окремим файлом PSPP виведе результати обчислення, про що повідомить мерехтіння значка PSPP (рис. 13).

Регресійний аналіз полягає у наближенні досліджуваного ряду розподілу результативного показника до ряду, який приблизно описує відповідність між результативною та факторними ознаками при цьому завдяки наближенню за можливості виключається дія випадкових факторів. Послідовність проходження регресійного аналізу така:

1. *Побудова емпіричної регресії.* На даному етапі передбачається побудова емпіричної лінії регресії на основі групування (аналітичного або комбінаційного) досліджуваної сукупності за факторною ознакою з визначенням середньозважених значень у кожній групі.

Емпірична регресія не передбачає виведення аналітичного виразу, що описує відповідність значень ознак, у результаті чого прогнозування ускладнено. Тут можливі два випадки:

1) Якщо значення факторної ознаки буде знаходитися в межах досліджуваного діапазону або на «розумній» відстані від нього, то вона буде належати до тієї чи іншої групи значень, якій буде відповідати середнє значення результативної ознаки. Але такі середні значення часто дуже грубо апроксимують дійсні значення результативної, оскільки для отримання статистично стійких груп необхідна досить велика кількість спостережень, що належать широкому діапазону значень, тому групові середні досить далеко знаходяться одна від одної. У такому разі для отримання

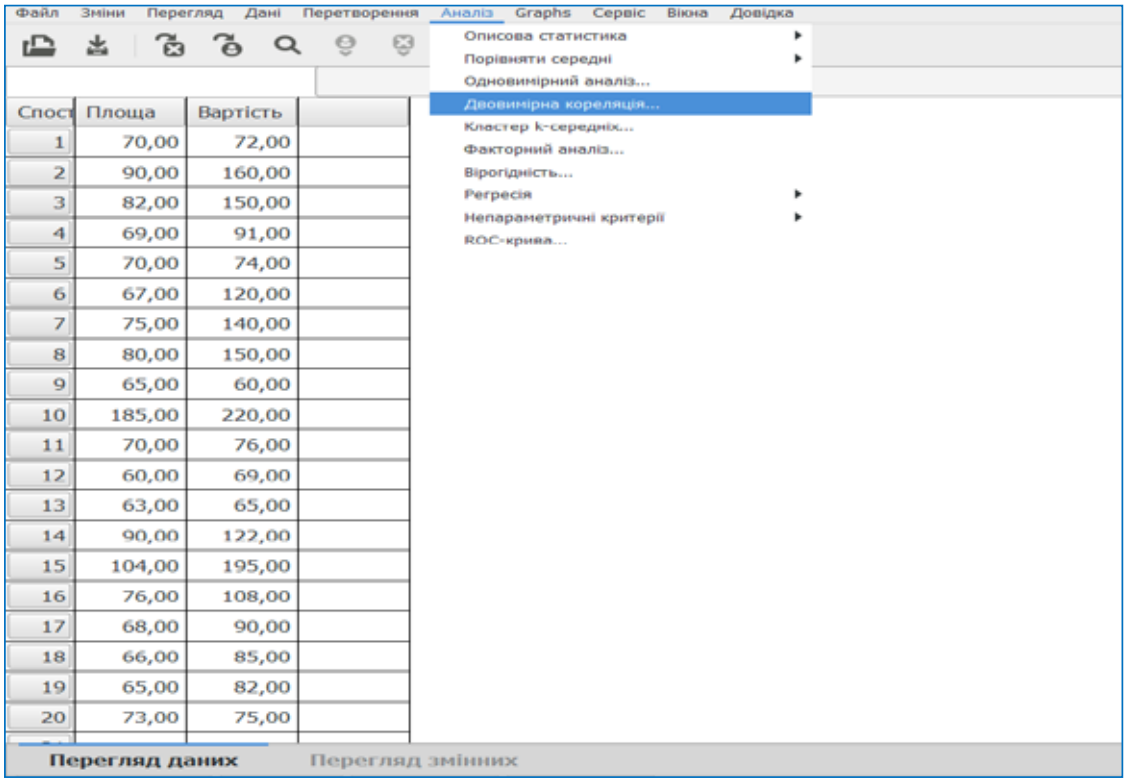

Рис. 10. Визначення парного коефіцієнта кореляції в PSPP

Джерело: створено автором

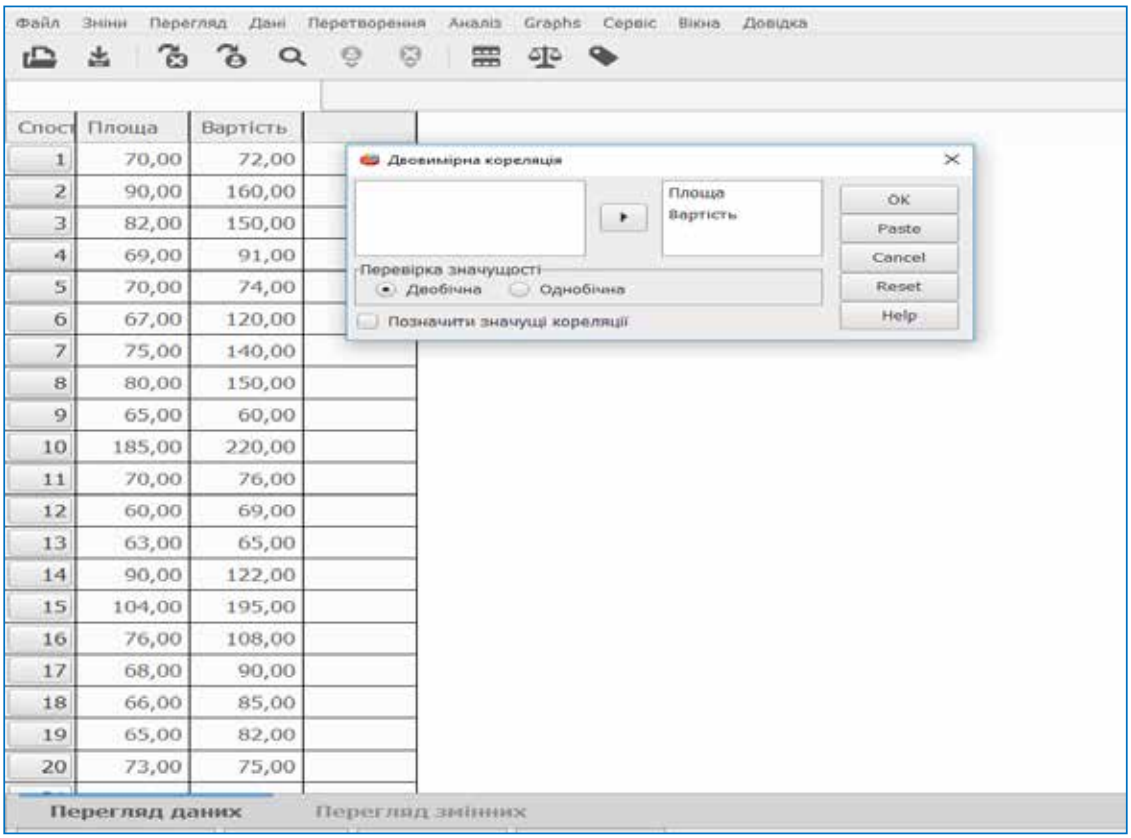

Рис. 11. Заповнення віконечка PSPP для визначення парного коефіцієнта кореляції

Джерело: створено автором

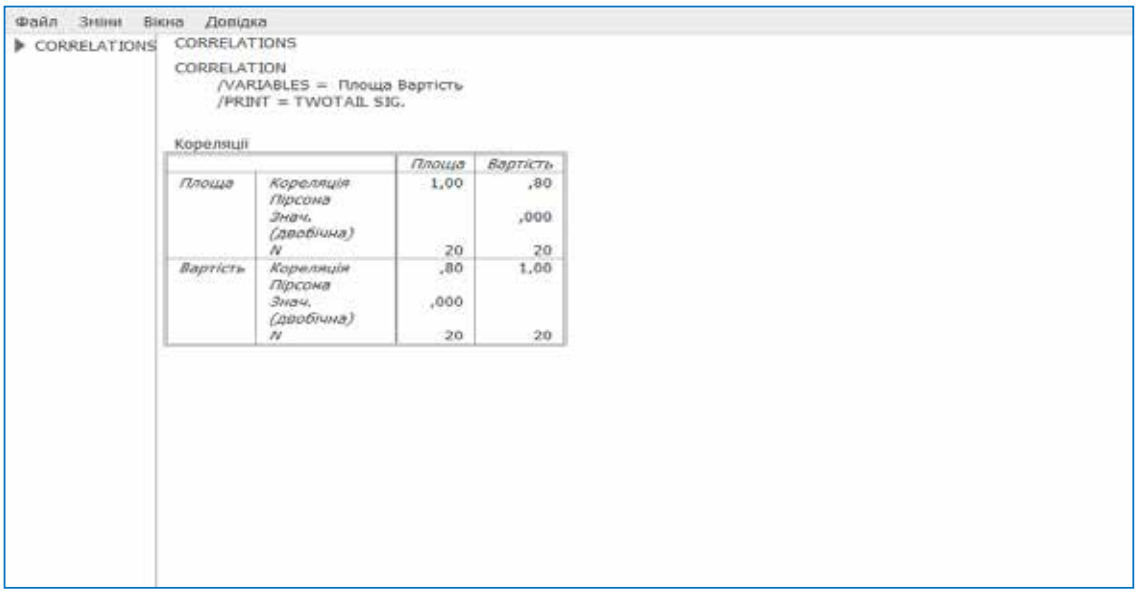

Рис. 12. Результати визначення щільності зв'язку між ознаками

Джерело: створено автором

прогнозного значення результативної ознаки необхідним є застосування методів екстраполяції або інтерполяції.

2) Значення факторної ознаки є «викидом» або потрібно спрогнозувати значення результативної ознаки на досить віддалений момент часу. Це завдання із застосуванням тільки емпіричної регресії не вирішити.

У силу зазначених причин побудова емпіричної регресії дає змогу виявити наявність і спрямованість та приблизно визначити її вид. Далі слід провести розрахунок параметрів залежно від методів аналітичної регресії.

2. *Побудова аналітичної регресії.* Побудова аналітичної регресії передбачає виведення функціональної залежності між факторними та результативною ознаками.

Першим етапом побудови аналітичної регресії є вибір виду залежності (специфікації моделі). На другому етапі визначають параметри залежності для вибраної специфікації моделі.

Необхідно відзначити, що вибір виду функціональної залежності є настільки ж важливим, як і найменш теоретично обґрунтованим. Вибір виду аналітичної залежності здійснюється з використанням одного з методів:

– графічного – на основі візуального аналізу кореляційного поля (графіка розсіювання даних);

– аналітичного – на основі аналізу накопиченої інформації про досліджуване явище;

– експериментального – методом перебору різних моделей і порівняння таких їх якісних показників, як залишкова дисперсія та середня помилка апроксимації.

Залежно від виду специфікації моделі розрізняють лінійну регресію і нелінійні регресійні моделі. Лінійна регресійна модель зображується поліномом першого ступеня. Нелінійна може проявлятися як у відношенні пояснюючих змінних, так і щодо оцінюваних параметрів.

Параметри залежності найчастіше оцінюють на підставі вибіркових даних із застосуванням методу найменших квадратів, який полягає у мінімізації функції суми квадратів різниці між функціональними й емпіричними значеннями залежної змінної (залишків).

Залежно від того, скільки факторних ознак включають у дослідження, розрізняють парну і множинну регресію (в першому випадку розглядається одна факторна ознака, у другому їх може бути декілька), а від виду функціональної залежності, до якої наближають досліджуваний ряд – лінійну і нелінійні регресії.

У разі парної лінійної регресії залежність між результативною та факторною ознаками (обидві з яких представлені в метричній шкалі) в генеральній сукупності представляють у вигляді лінійної функції.

На основі вибіркових даних оцінюються параметри вибіркового рівняння регресії. Розрахувавши параметри рівняння регресії, слід оцінити його якість, тобто ступінь, в якій реальні (спостережені) значення результативної ознаки відповідають теоретичним, розрахованим за аналітичним виразом. Таким чином, можна оцінити прогнозну силу моделі, що вийшла, тобто наскільки за поведінкою факторної ознаки на основі даної моделі можна оцінити поведінку результативної ознаки. Така оцінка базується на розкладанні загальної варіації залежної змінної на два складника: варіацію, зумовлену поведінкою (зміною) факторної ознаки та випадкову варіацію. Ці варіації оцінюються за допомогою таких

показників, як загальна дисперсія, факторна дисперсія і залишкова дисперсія. У реальній ситуації ці параметри невідомі, але їх можна оцінити на основі аналогічних вибіркових показників.

Незміщені оцінки генеральних параметрів знаходять шляхом ділення відповідної суми квадратів відхилень на число ступенів свободи цього виразу.

Оцінка варіації залежної змінної здійснюється за допомогою таких показників, як середня помилка апроксимації, або модуль середнього лінійного відхилення, та коефіцієнт детермінації, який показує частку дисперсії, що пояснюється дією факторної ознаки. У разі парної лінійної регресії коефіцієнтом детермінації є квадрат коефіцієнта кореляції. При цьому слід урахувати, що розрахунок коефіцієнта детермінації коректний, якщо в рівняння регресії включена константа.

Оскільки коефіцієнт детермінації розраховується на підставі тільки вибіркових даних, то необхідно оцінити його значимість для всієї сукупності, тобто розрахувати рівень його значущості на підставі даних про обсяг вибірки. Значимість коефіцієнта детермінації оцінюється за допомогою F-критерію Фішера.

Для сукупностей з обсягом n <30, окрім аналізу варіації залежної змінної, проводять аналіз варіації параметрів регресії, тобто оцінюють, наскільки їх вибіркові оцінки відхиляються від дійсних значень і наскільки такі оцінки значущі. Дійсно, оцінки параметрів регресії, так само як і величину результативної ознаки в моделі, можна представити у вигляді суми двох складників – випадкового і невипадкового. Невипадковий буде відповідати дійсному значенню параметра, випадковий – відхиленню від цього дійсного значення.

На основі оцінок стандартних відхилень параметрів регресії та їхніх індивідуальних значень можна перевіряти гіпотези рівності коефіцієнтів регресії заданим величинам. Для цього визначається значення t-критерію, що розраховується як відношення різниці наявної оцінки коефіцієнта і заданої величини до оцінки стандартного відхилення коефіцієнта.

Якщо отримана величина t-критерію більша за критичне його значення для заданого рівня значущості, то різниця є значущою і гіпотеза рівності даного параметра регресії заданої величиною відхиляється.

Для визначення показників регресійної статистики в PSPP необхідно виконати такі кроки: «Аналіз» → «Регресія» → «Лінійна» (рис. 13).

Заповнюємо віконечко (рис. 14).

Результати будуть надані окремим файлом результатів (рис. 15).

Заключним етапом усіх статистичних розрахунків є аналіз отриманих результатів.

Етап аналізу отриманої інформації − це зіставлення отриманої про вивчений об’єкт інформації з уже відомим об’ємом знань про нього.

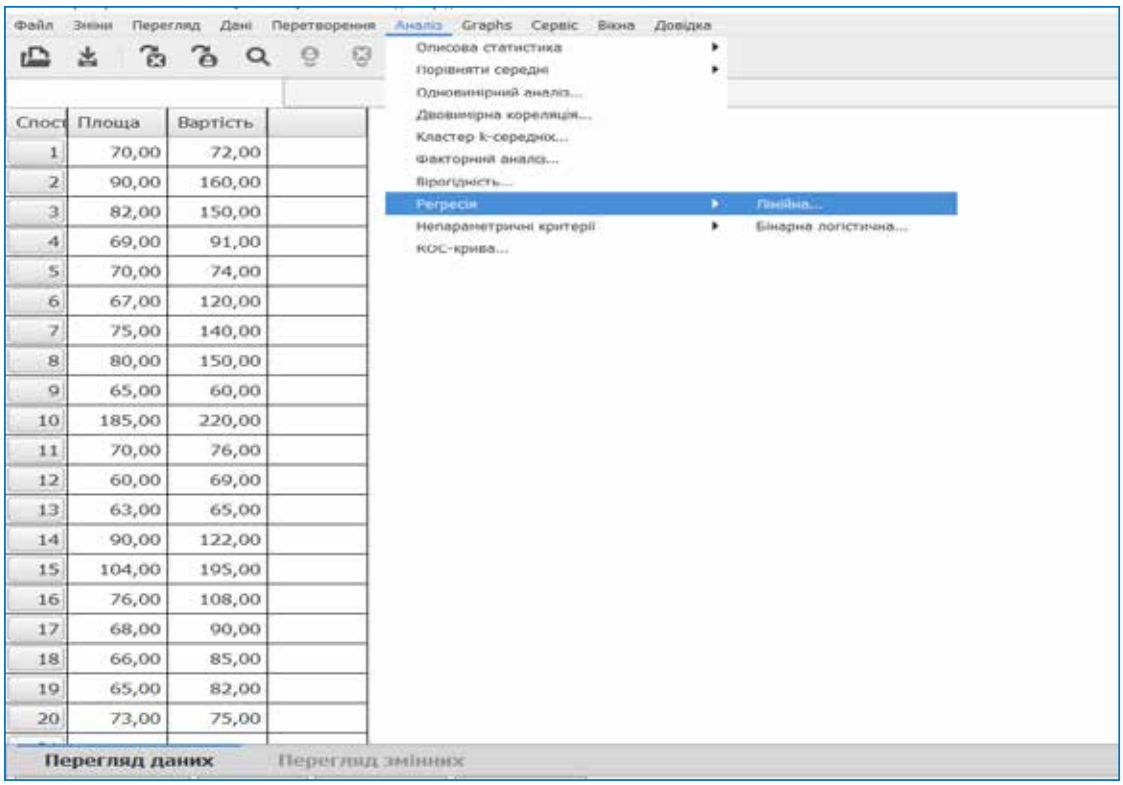

Рис. 13. Визначення лінійної регресії в PSPP

Джерело: створено автором

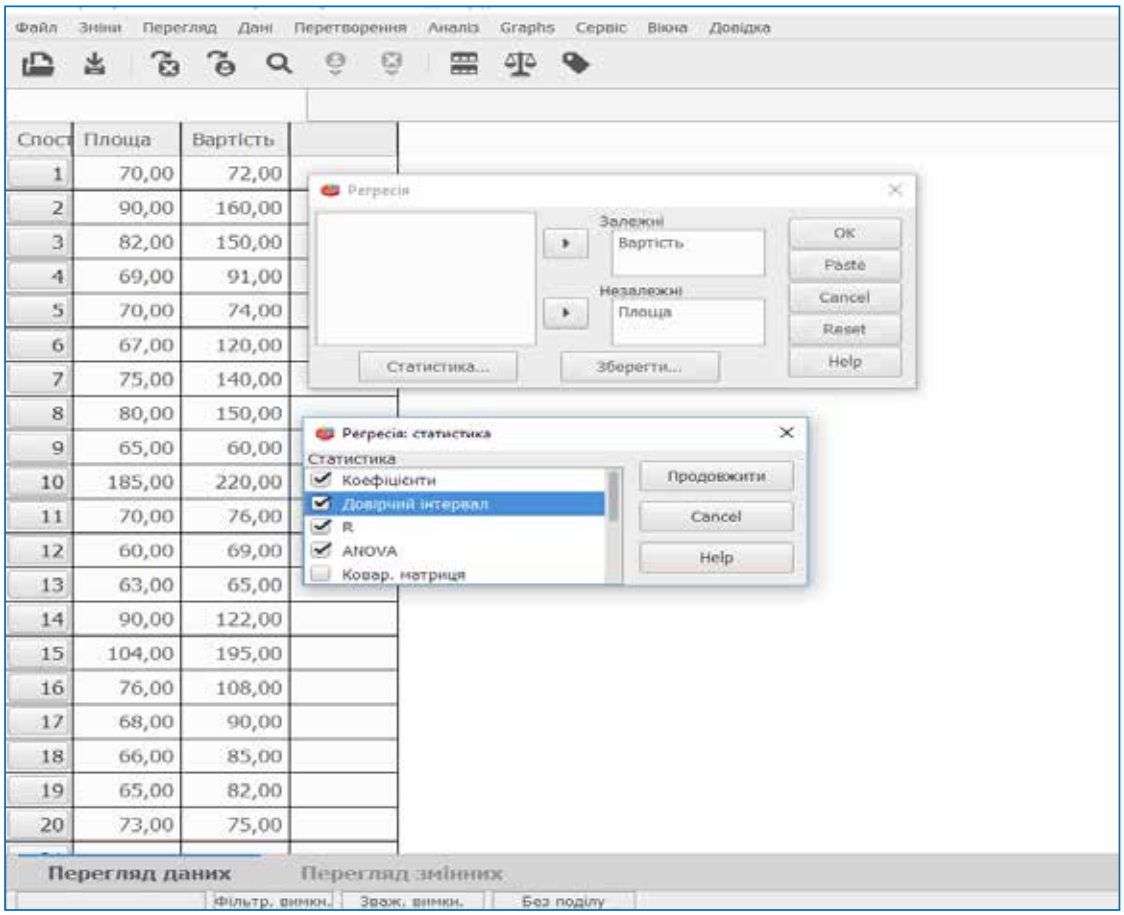

Рис. 14. Обчислення показників регресії в PSPP

Джерело: створено автором

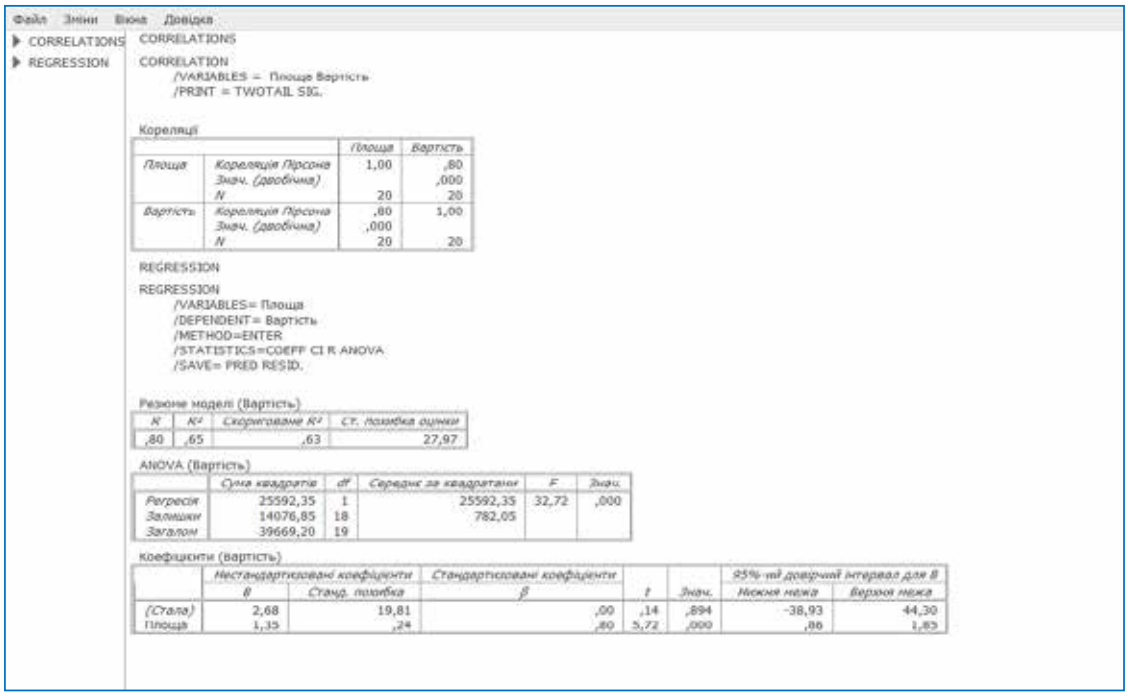

Рис. 15. Результати регресійного аналізу в PSPP

Джерело: створено автором

Основною метою аналізу даних дослідження є пояснення змісту окремих результатів, об'єднання і виділення узагальнюючих положень, зведення їх в одну теоретичну систему.

Для отримання надійних і достовірних результатів емпіричного аналізу варто дотримуватися низки вимог:

1) дослідник повинен мати уявлення про логіку використаних статистичних методів;

2) застосовувати правильно підібрані статистичні методи для аналізу даних досліджень, адже множина статистичних методів буває достатньо різноманітною залежно від типу дослідження: емпіричного, прикладного або теоретичного.

3) варто попередньо провести пробну обробку на невеликій кількості масиву даних.

У процесі аналізу та узагальнення результатів умовно можна виділити кілька етапів:

1) упорядкування, класифікації, групування даних згідно з дослідницькими гіпотезами;

2) узагальнення даних, перевірку значущості й достовірності числових характеристик;

3) перевірку дослідницьких гіпотез за допомогою отриманих числових характеристик.

Висновки з цього дослідження і перспективи подальших розвідок у даномунапрямку. Отже, для проведення глибокого та ефективного статистичного аналізу доцільно використовувати методи описової статистики, кореляційного та регресійного аналізу. Швидко та ефективно обробити інформацію за даними напрямами аналізу можна за допомогою універсальної програми PSPP, яка має широкий спектр можливостей статистичної обробки даних, не потребує коштів на придбання та обслуговування, а також глибокої математичної підготовки користувачів.

БІБЛІОГРАФІЧНИЙ СПИСОК:

1. Огляд програмних засобів статистичного аналізу даних / М.В. Роїк, О.І. Присяжнюк, В.О. Денисюк. Ефективна економіка. 2017. № 7. URL: http://www.m.nayka.com.ua/?op=1&j=efektyvna-ekonomika&s=ua&z=5676.

2. Айвазян С.А., Степанов В.С. Программное обеспечение по статистическому анализу данных: методология сравнительного анализа и выборочный обзор рынка. URL: http://pubhealth.spb.ru/SAS/STatProg.htm.

3. Василенко Ж.В. Программное обеспечение по статистическому анализу данных. Методология сравнительного анализа. URL: http://www.giac.unibel.by/sm_full.aspx?guid =8313.

4. Майборода Р.Є., Сугакова О.В. Статистичний аналіз даних за допомогою пакету STATISTICA. URL: http://matphys.rpd.univ.kiev.ua/downloads/courses/mmatstat/StatAn.doc.