Л.Л. Маханець, О.Ю. Вінничук, М.В. Григорків

Статистика:

лабораторний практикум у STATISTICA 12

ЧНУ РУТА
ISBN 978-966-423-763-2
9 789664 237632

Міністерство освіти і науки України
Чернівецький національний університет
імені Юрія Федьковича

Л.Л. Маханець, О.Ю. Вінничук, М.В. Григорків

СТАТИСТИКА:

ЛАБОРАТОРНИЙ ПРАКТИКУМ У STATISTICA 12

Навчальний посібник

Чернівці
Чернівецький національний університет
імені Юрія Федьковича
2023

УДК 311(076.5)
М 360

Друкується за ухвалою Вченої ради
Чернівецького національного університету імені Юрія Федьковича
(протокол № 13 від 28.12.2021 року)

Рецензенти:

Вітлінський В.В. – доктор економічних наук, професор, професор кафедри математичного моделювання та статистики ДВНЗ "КНЕУ імені Вадима Гетьмана"

Вдовічен А.А. – доктор економічних наук, професор, директор Чернівецького торговельно-економічного інституту КНТЕУ, професор кафедри менеджменту, міжнародної економіки та туризму

Маханець Л.Л., Вінничук О.Ю., Григорків М.В.

М 360 Статистика: лабораторний практикум у STATISTICA 12: навч. посіб. Чернівці : Чернівец. нац. ун-т ім. Ю. Федьковича, 2023. 161 с.

ISBN 978-966-423-763-2

У навчальному посібнику наведено завдання до лабораторних робіт із статистики та приклади їх виконання у системі статистичного аналізу STATISTICA 12. Кожна лабораторна робота доповнена відповідними завданнями для самостійної роботи і спрямована на засвоєння практичного інструментарію обробки статистичної інформації.

Для студентів вузів економічних напрямів підготовки.

УДК 311(076.5)

ISBN 978-966-423-763-2

Навчальне видання

Маханець Любов Леонідівна
Вінничук Олена Юріївна
Григорків Марія Василівна

СТАТИСТИКА:

ЛАБОРАТОРНИЙ ПРАКТИКУМ У STATISTICA 12

Навчальний посібник

Відповідальний за випуск *Григорків В.С.*
Літературний редактор *Лупул О.В.*
Технічний редактор *Кудрінська О.М.*

Підписано до друку 03.02.2023. Формат 60 х 84/8.
Папір офсетний. Друк різографічний. Ум.-друк. арк. 18,6 .
Обл.-вид. арк. 20,0. Зам. Н-005.
Видавництво та друкарня Чернівецького національного університету 58002,
Чернівці, вул. Коцюбинського, 2

Свідоцтво суб'єкта видавничої справи ДК №891 від 08.04.2002 р.

ЗМІСТ

ПЕРЕДМОВА

Статистика є нормативною навчальною дисципліною, з вивченням якої починають формуватися необхідні професійні навички економістів, менеджерів, підприємців, маркетологів, аналітиків тощо. Оволодіння статистичною методологією є однією з неодмінних умов прийняття оптимальних рішень на всіх рівнях підприємницької діяльності. У результаті вивчення даної дисципліни студенти набувають навичок збору, обробки та аналізу статистичної інформації, виявлення та оцінки закономірності розвитку та взаємодії складних соціально-економічних явищ. Важливу роль при цьому відіграє формування навичок виконання розрахункових операцій, зокрема з використанням комп'ютерної техніки та новітнього програмного забезпечення.

Одним з безперечних лідерів серед спеціалізованих додатків для статистичного аналізу є STATISTICA, яка розроблена американською фірмою StatSoft, Inc. в 1991 році. На сьогодні існує вже 14 версія системи STATISTICA. У навчальному посібнику наведено приклади виконання практичних завдань у системі статистичного аналізу STATISTICA 12, оскільки нововведення 14 версії не значно змінили саме статистичний аналіз. Вони стосуються баз даних, Data Mining та інтеграції R й Python.

STATISTICA повністю узгоджена зі стандартами Windows, її досить легко опанувати (завдяки відображення інтуїтивних уявлень статистиків про середовище аналізу даних), мінімальні технічні вимоги до комп'ютера, унікальна презентаційна і наукова графіка, вичерпний набір сучасних та класичних методів статистики, поданих у системі.

STATISTICA – це система для статистичного аналізу даних, що містить широкий набір аналітичних процедур і методів: більше 100 різних типів графіків, описові та внутрішньогрупові статистики, статистичний аналіз даних, таблиці частот та спряженості, лінійна та множинна регресія, непараметричні статистики, загальна модель дисперсійного і коваріаційного аналізу, підгонка розподілів і багато інших видів статистичного аналізу.

Продукти серії STATISTICA основані на найсучасніших технологіях, повністю відповідають останнім досягненням в області інформаційних технологій, дозволяють вирішувати будь-які завдання в області аналізу і обробки даних, ідеально підходять для вирішення практичних завдань маркетингу, фінансів, страхування, економіки, бізнесу, промисловості, медицини тощо.

У навчальному посібнику викладено основні етапи статистичного дослідження, що виконуються за допомогою системи STATISTICA. Всі лабораторні роботи відповідають навчальній програмі зі статистики. Практикум містить необхідні матеріали, які сприятимуть самостійному вивченню навчальної дисципліни, а також методичні вказівки до виконання лабораторних робіт з відповідних тем навчальної дисципліни та завдання для самостійної роботи.

Оскільки програма STATISTICA є продуктом американської компанії, вона має англомовний інтерфейс.

Посібник призначений для студентів вищих навчальних закладів, аспірантів, викладачів, науковців і практиків, діяльність яких пов'язана із застосуванням методів статистичного аналізу даних.

Лабораторна робота № 1
Початок роботи із системою STATISTICA 12

1. Запуск системи STATICTICA 12

Запуск системи **STATISTICA** здійснюється аналогічно запуску будь-якого Windows-додатку. Вікно системи схоже на вікна інших додатків і складається з наступних основних елементів: рядка заголовків, рядка меню, панелі інструментів, робочої зони та рядка стану.

STATISTICA має модульну організацію, тобто система має кілька модулів, кожен з яких працює автономно і має можливості для обміну даними з іншими модулями системи.

При першому відкритті система **STATISTICA 12** запропонує вибрати зручний для користувача інтерфейс (рис. 1.1). Вибрана панель інструментів буде використовуватись для подальшої роботи зі всіма файлами системи.

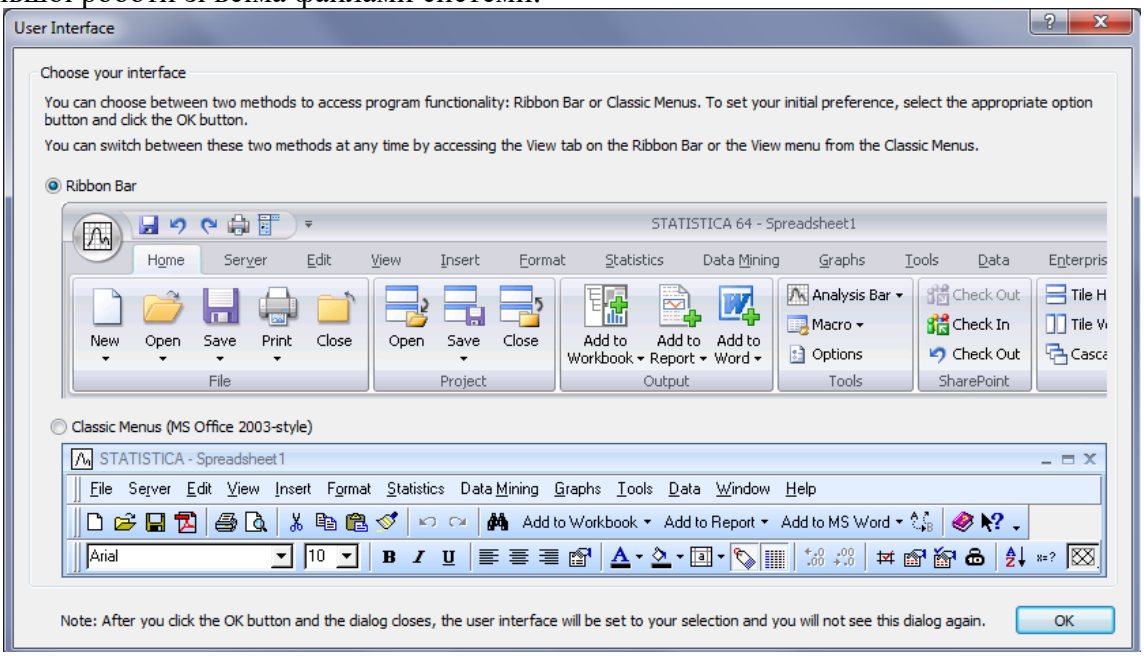
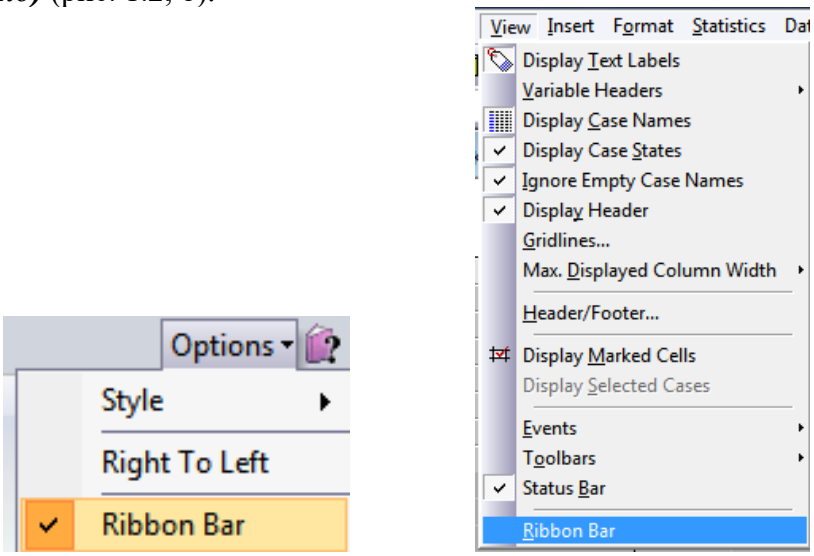


Рис. 1.1. Вибір інтерфейсу при першому завантаженні системи

Якщо вибрана панель інструментів не влаштовує, то можна переключитися. У разі переходу до класичного меню від стрічки, то необхідно зняти прапорець ***Ribbon Bar (Стрічка)*** з меню ***Options (Опції)***, для вибору класичної панелі інструментів (рис. 1.2, а); для переходу до стрічки з класичного інтерфейсу необхідно вибрати ***Ribbon Bar (Стрічка)*** з меню ***View (Вигляд)*** (рис. 1.2, б).



а) б)

Рис. 1.2. Вибір панелі інструментів

Після відкриття **STATISTICA** завантажується новий файл або файл даних, з яким працювали попередньо. Одночасно **STATISTICA** відкриває діалогове вікно початку роботи, в якому можна обрати одну з наступних дій: відкрити файл даних системи **STATISTICA**, книгу MS Excel, зробити запит до зовнішньої бази даних, відкрити звіт, робочу книгу, макрос, R–сценарій, проект Data Miner, проект системи STATISTICA, звернутися до електронного підручника, переглянути відео з описом функціональних можливостей системи, список файлів, що недавно використовувалися або вибрати опцію ***Don't show this dialog again (Не показувати цей діалог знову)***, щоб він не з'являвся в подальшій роботі в системі. Можна також відмовитися виконувати будь-яку із запропонованих дій і закрити діалогове вікно натисканням кнопки ***Close (Закрити)*** (рис. 1.3).

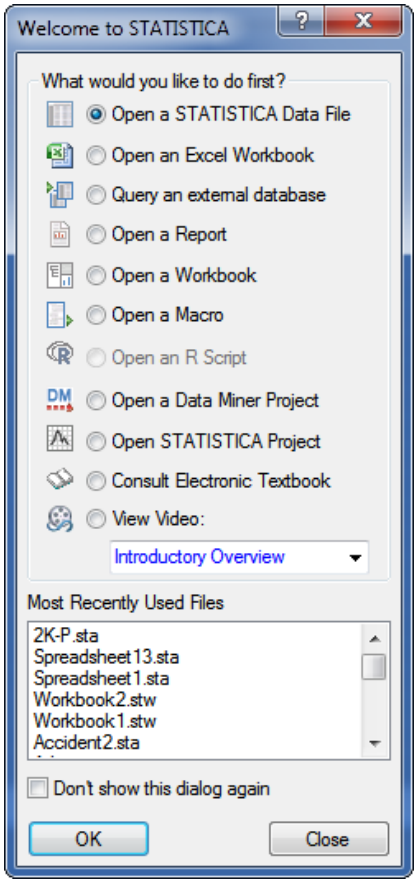


Рис. 1.3. Діалогове вікно початку роботи

Дані в **STATISTICA** подаються у вигляді електронної таблиці (рис. 1.4). Таблиця з вхідними даними (таблиці зберігаються у файлах з розширенням *.sta) є одним з типів файлу в системі **STATISTICA** (інші типи файлів – електронна таблиця з результатами аналізу, графік, звіт тощо). Кожен тип файлу виводиться в своєму вікні в робочій області системи і зберігається як окремий файл. Як тільки вікно відповідного файлу стає активним, змінюється панель інструментів і меню. У ньому з'являються команди, доступні для цього типу файлів.

На відміну від звичайних електронних таблиць (наприклад, в MS Excel), де рядки і стовпці рівноправні, в **STATISTICA** вони мають різні змістові значення. При цьому стовпці таблиці називаються ***Variables (Змінні)***, а рядки – ***Cases (Спостереження)***. Кожна змінна має своє ім'я, формат та інші атрибути, які називаються специфікацією змінної та задаються користувачем. Електронна таблиця з вихідними даними в **STATISTICA** називається ***Spreadsheet***.

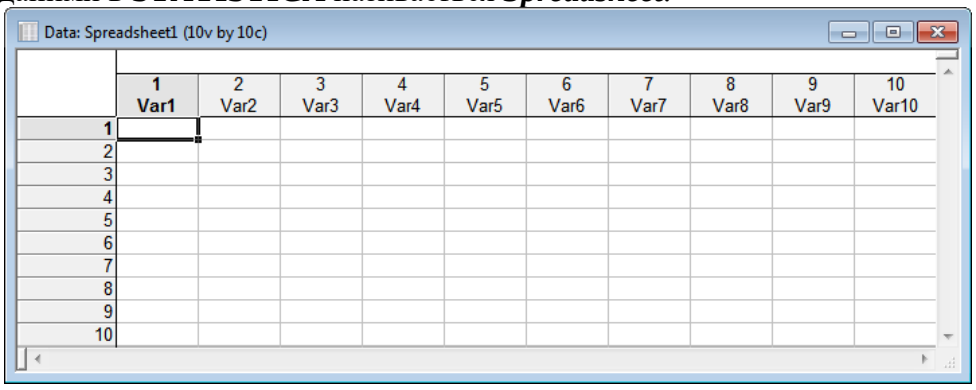


Рис. 1.4. Вигляд електронної таблиці для введення даних

2. Створення нового файлу

Створення нового файлу з даними в системі **STATISTICA** може бути здійснено за допомогою групи ***File (Файл)*** вкладки ***Home (Основне)*** або меню ***File (Файл)***[1], вибравши команду ***New*** (рис. 1.5). Відкриється діалогове вікно створення файлу, що містить наступні

[1] Порядок дій вибору потрібної команди в системі Statistics описується спочатку у режимі ***Ribbon Bar***, а потім – у ***Classic Menus***.

вкладки: таблиця, звіт, макрос, робоча книга, база даних, браузер та документи Office, параметри яких можна використовувати для створення нових файлів.

Для створеної нової електронної таблиці потрібно у діалоговому вікні (рис. 1.6) виділити вкладку ***Spreadsheet (Таблиця)*** і вказати ***Number of variables (Число змінних)*** і ***Number of cases (Число спостережень***). Більшість процедур **STATISTICA** працюють до розмірності 300 змінних, хоча загальна кількість стовпчиків, що зберігаються у файлі, може досягати кількох тисяч. **STATISTICA** дозволяє зберігати 99999 (100000–1) рядків.

Також можна вказати ***Case name length (Довжину імені спостереження)***, ***MD code (Код для пропущених даних)***, ***Default data type (Тип даних за замовчуванням)***, ***Placement (Розміщення)***, ***Var name prefix (Префікс імені змінної)***, ***Var name start number (Стартовий номер змінної)***, ***Display format (Формат відображення змінних).*** Щоб відмінити всі налаштування, треба натиснути кнопку ***Default (За замовчуванням).*** При встановленні всіх налаштування вибирають ***OK*** (рис. 1.6).

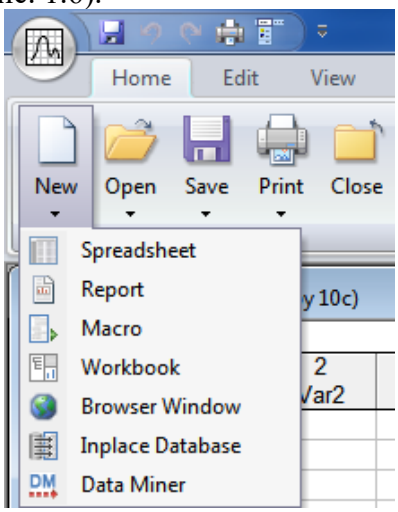


Рис. 1.5. Створення нового файлу за допомогою кнопки ***New*** (вкладка ***Home*** групи ***File***)

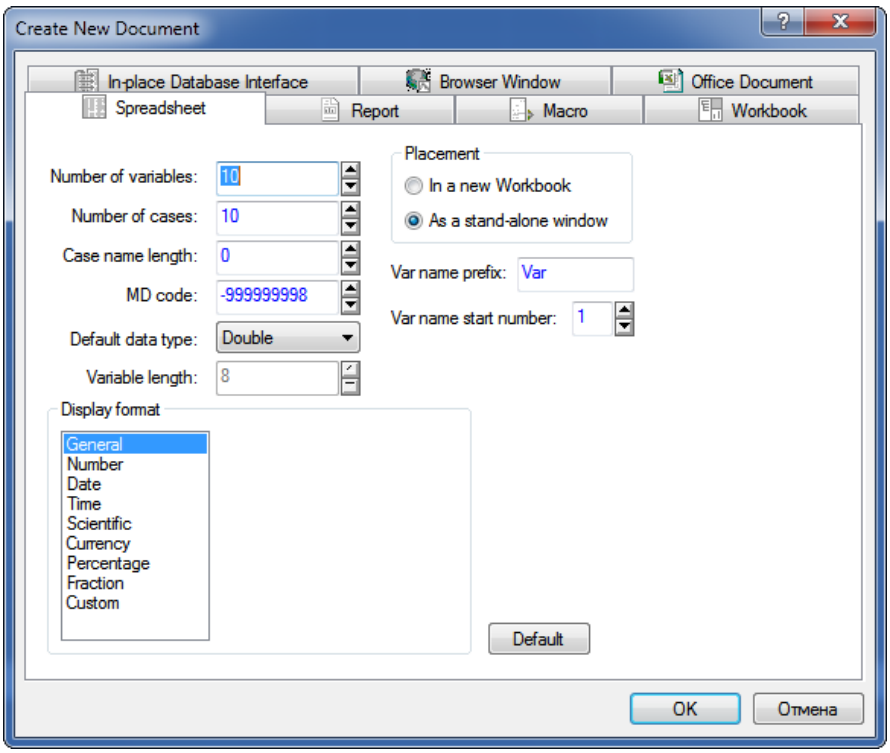


Рис. 1.6. Діалогове вікно ***Create New Document***

3. Параметри змінних

Для задання імен та параметрів змінних можна двічі клацнути на полі введення імені

змінної і вказати необхідне ім'я або встановити параметри. Інший варіант задання імені або параметрів змінної можливий вибравши вкладку ***Data (Дані)***→групу ***Variables*** (рис. 1.7) або за допомогою кнопки ***Vars (Змінні)*** на панелі інструментів ***Spreadsheet*** чи через контекстне меню, клацнувши правою кнопкою миші на імені змінної та вибравши в контекстному меню ***Variable Specs… (Специфікація змінної)***.

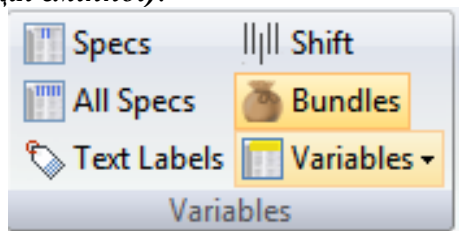


Рис. 1.7. Група ***Variables***

Група ***Variables*** вкладки ***Data*** містить такі кнопки (рис. 1.7):

- ***Specs***;
- ***All Specs***;
- ***Text Labels***;
- ***Shift***;
- ***Bundles***;
- ***Variables***.

Опишемо зміст деяких кнопок групи ***Variables***.

Кнопка ***Specs*** групи ***Variables*** дозволяє задати параметри поточної змінної. У діалоговому вікні ***Variable*** (рис. 1.8) можна задати:

- шрифт імені змінної;
- ***Name (Ім'я)*** – ім'я змінної;
- ***Type (Тип)*** – тип змінної: ***Double (Подвійної точності)*** – діапазон значень, що підтримується цим типом даних становить приблизно ± $1{,}7 \cdot 10^{308}$; ***Integer (Ціле)*** – діапазон від –2147483648 до 2147483648; ***Text (Текст)***, ***Byte (Байт)*** – діапазон від 0 до 255;
- ***Measurement Type (Тип вимірювань)*** – шкалу вимірювання змінної: ***Auto (Авто)***, ***Unspecified (Будь-яка)***, ***Continuous (Неперервна)***, ***Categorical (Категоріальна)*** чи ***Ordinal (Порядкова)***;
- ***Length (Довжина)*** – довжину текстової змінної;
- ***Excluded (Виключено)*** – виключення змінної з аналізу/графіків.

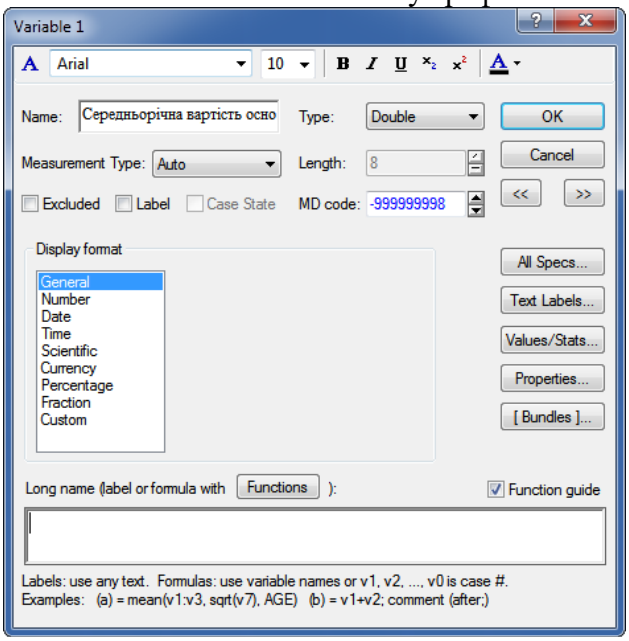


Рис. 1.8. Діалогове вікно специфікації змінної

- ***Label (Позначення)*** – використання значення обраної текстової змінної у відповідних графах.
- ***MD Code (Код для пропущених даних)*** – вказання кодів для пропущених змінних.
- ***Display format (Формат відображення змінних)*** – формат подання даних: ***General (Загальний)***, ***Number (Числовий)***, ***Date (Дата)***, ***Time (Час)***, ***Scientific (Науковий*** (наприклад, 1.2345E-02)***)***, ***Currency (Грошовий)***, ***Percentage (Відсотковий)***, ***Fraction (Дробовий)***, ***Custom (Користувацький*** (задається користувачем)***).***
- ***Long Name (label or formula) (Довге ім'я (мітка або формула))*** – додаткову інформацію про змінну або формулу для обчислень.
- ***All Specs…*** – параметри для всіх змінних.
- ***Text Labels…*** – редагування текстових значень. У багатьох випадках **STATISTICA** автоматично кодує текстові змінні (наприклад, стать – чоловіча чи жіноча, списки міст, професій тощо), призначивши їм цілі порядкові номери, починаючи зі 101. Проте можна вносити корективи власноруч, використовуючи диспетчер текстових значень, який відображає змінну, на якій стоїть курсор (рис. 1.9). Потім можна вносити тільки коди текстових значень, що значно полегшить набір вхідних даних.

Text Label	Numeric	Description
жіноча	101	
чоловіча	102	

Рис. 1.9. Кодування текстових змінних

- ***Values/Stats…*** – показує дані по змінній (назву, набір значень, середнє тощо).

У спадному списку кнопки ***Vars*** (***Variables*** у ***Ribbon Bar*** групи ***Variables***) можна вибрати такі функції:

- ***Add…*** – додати змінну. Максимальна кількість змінних – 4092. Модуль диспетчера великих файлів даних (Megafile Manager) може зберігати файли даних із 32000 змінними.
- ***Move…*** – перемістити змінну. Переміщує стовпець даних на вказану позицію. Можна вказати список змінних, які потрібно перемістити. У меню, що з'являються треба вказати такі параметри:
  - ✓ ***From Variable, To Variable –*** початковий і кінцевий номер (ім'я) змінних, які потрібно перемістити.
  - ✓ ***Insert After –*** номер (ім'я) змінної, після якої потрібно вставити змінні.
- ***Copy…*** – скопіювати на зазначене місце стовпці з їх вмістом. У діалоговому вікні команди необхідно задати параметри: з якої і по яку змінну скопіювати; після якої змінної вставити.
- ***Delete…*** – видалити стовпці. У діалоговому вікні команди треба вказати імена змінних початку і кінця діапазону видалення.

Кнопка ***Shift (Зсув)*** групи ***Variables*** призначена для переміщення однієї чи діапазону змінних на потрібну кількість рядків (***Forward (уперед)*** чи ***Backward (назад)***). У діалоговому вікні вибирають змінні зі списку чи рядки спостереження, вказують потрібні параметри: ***Lag*** – визначення кількості рядків, на які потрібно перемістити змінну; ***Direction*** – визначення напряму переміщення змінної (уперед чи назад).

4. ***Керування даними***

У системі **STATISTICA** є спеціальні можливості керування даними (вкладка ***Data***), які дозволяють швидко створити електронну таблицю, об'єднати дві таблиці, вирізати частину таблиці, відсортувати спостереження за якою-небудь ознакою тощо. Розглянемо деякі групи вкладки ***Data***.

Група ***Transformations (Перетворення)*** містить такі можливості (рис. 1.10).

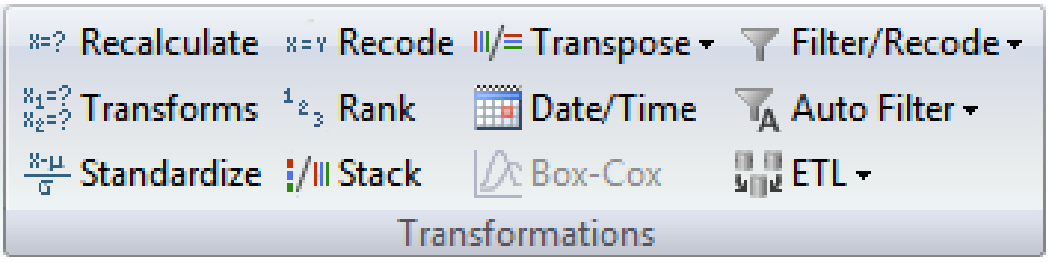


Рис. 1.10. Група ***Transformations*** вкладки ***Data***

- ***Recalculate (Перерахунок)*** – обчислення за формулами. Система робить обчислення тільки за запитом і попереднє значення не відновлюється. Проте в цьому є і свої переваги: не потрібно створювати зайві комірки, в які записуватиметься результат розрахунку формул. Результат обчислення буде відображатися в тих комірках, в яких були дані для обчислення. Наприклад, необхідно визначити квадратний корінь із значення змінної номер 6 (у формулах змінна задається як v+номер (приклад, v6) незалежно від назви). У полі ***Long Name (label or formula)*** змінної номер 6 записуємо =Sqrt(v6) (рис. 1.8), після чого запускаємо ***Recalculate***.
- ***Transforms*** – перетворення змінної відповідно до введеної формули, однак розрахунки можна записати в іншому стовпці.
- ***Standardize (Стандартизація даних)*** – стандартизація даних зведенням середнього до нуля, а дисперсії – до одиниці.
- ***Recode (Перекодування змінних)*** – перекодування вихідних значень поміченої змінної у нові значення. Меню діалогу ***Recode Value of Variable*** має вигляд:
  - ✓ ***Category (Категорія)***. Якщо вмикати вибір за умовою (***Include/Exclude if***), то попереднє значення буде замінене новим лише тоді, коли умова виконуватиметься. Задання умов перекодування збігається зі стандартним алгоритмом вибору за умовою. Використовуються оператори =, <>, <, >, <= >=, NOT, AND, OR. Імена змінних вказуються в їх короткій формі (наприклад, AGE) чи у відповідності з їх порядковим номером (v1, v2, v3,...). Вказівка на рядки дається символом v0 (якщо потрібно виділити перші 100 рядків із файлу даних, то це буде записано так: v0<101). Так можна визначати дуже складні конструкції вибору за умовою. Усі посилання на текстові значення даних беруться в одиночні лапки (’Yes’).
  - ✓ ***New Value*** – визначає нове значення, яке замінюватиме старі при перекодуванні (Value). Можна також призначити вибраній категорії даних код пропущених значень (MD code).
- ***Rank (Ранжування змінних)*** – значення даних у вибраній змінній будуть замінені значеннями рангів відповідно до:
  - ✓ ***Assign Rank 1 to –*** присвоєння рангу 1 найменшому чи найбільшому значенню змінної.
  - ✓ ***Ranks for Ties –*** ранжування пов’язаних значень (залежних):
    - ***Mean*** (присвоює ранг пов’язаним значенням як середнє цих рангів);
    - ***Sequential*** (послідовно ранжує кожне значення, не зважаючи на зв’язки);
    - ***Low*** (присвоює найменший ранг пов’язаним значенням);
    - ***High*** (присвоює найбільший ранг пов’язаним значенням);
  - ✓ ***Type of Ranks*** (тип ранжування):
    - ***Regular*** (регулярний, ранжує від 1 до n);
    - ***Fractional*** (фракціональний, ранжує від 0 до 1);
    - ***Fractional as %*** (фракціональний відсотковий, ранжує від 0 до 100%).

- ***Date/Time (Дата/час)*** – перетворення дат, які записані в одній змінній (наприклад, 01.09.1997) у дві чи три змінні, у кожній із яких стоятиме відповідна категорія – рік, місяць і день. Протилежна процедура також можлива – зведення дат дня, року, місяця в одне ціле.

У групі ***Transformations*** (рис. 1.10) є також можливість ***фільтрування даних (Filter/Recode)***, яка аналогічна відповідній функції в MS Excel.

Група ***Manage (Управління)*** вкладки ***Data*** містить такі кнопки:

- ***Merge (Злиття файлів)*** – об'єднання даних в один файл. Оскільки не завжди змінні й рядки мають однакові назви і співпадають, тому **STATISTICA** пропонує такі опції:
  - ✓ ***Merge Variables*** – об'єднання змінних;
  - ✓ ***Merge Cases*** – об'єднання рядків;
  - ✓ ***Text Labels*** – об'єднання текстових значень змінних.
- ***Subset (Вибір змінної за певних умов)*** – формування списку змінних згідно заданих умов (рис. 1.11).
- ***Verify Data (Верифікація змінних)*** – задання умов, за якими виконуватиметься перевірка (зазначаються так само, як спосіб утворення підмножин у файлі даних (***Case selection***) (рис. 1.12).

Меню верифікації даних (рис. 1.12) має такі налаштування: ***All conditions are met*** – виконані всі зазначені умови, або ***At least one condition is met*** – хоча б одна умова була виконана.

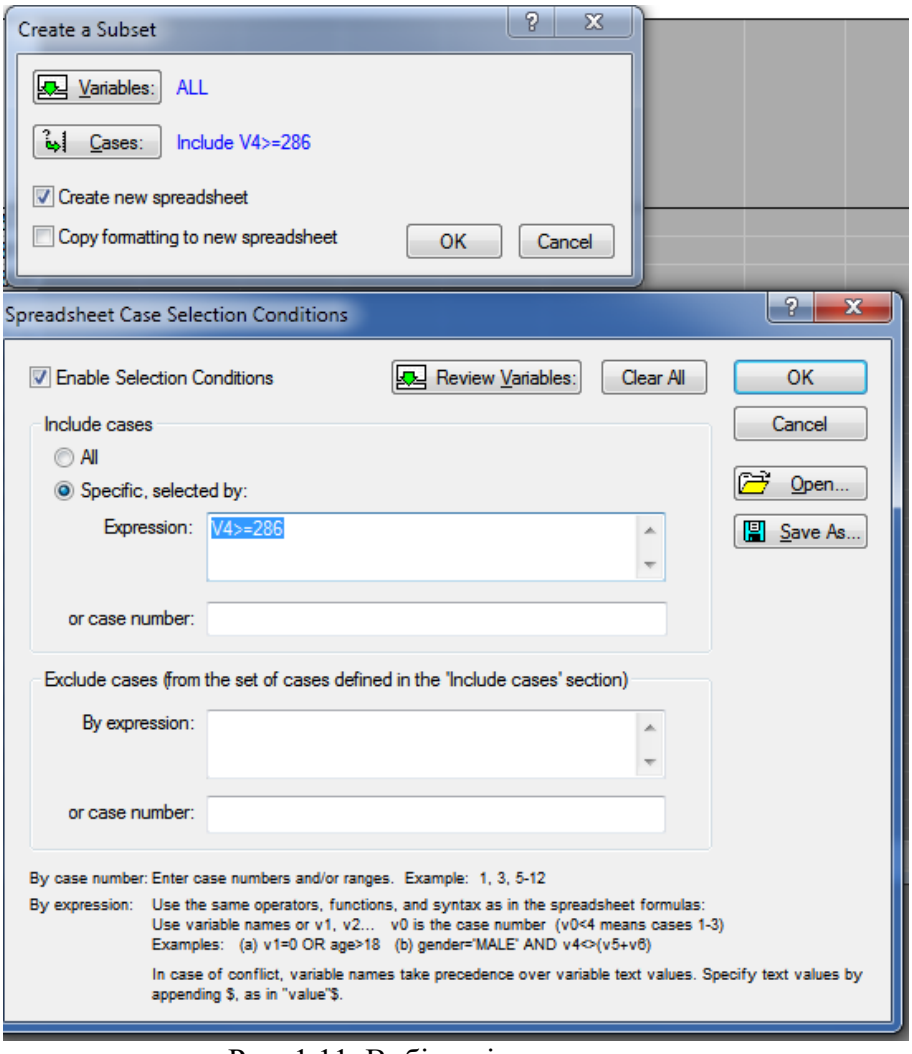


Рис. 1.11. Вибір змінних за певних умов

Зазначивши умови перевірки, потрібно вибрати кнопку ***Find First***, і **STATISTICA** почне перевіряти дані. Рядок, який не відповідає умовам перевірки, підсвічується. Якщо вибрати кнопку ***Mark All Invalid***, підсвітяться всі дані, що не відповідають умовам.

У полях ***Condition (Умова)*** (рис. 1.12) задаються умови. Потрібно тільки зауважити, яке значення умови перевірки задовольняє: ***Valid if (коректне)*** або ***Invalid if (помилкове)***.

Широкий набір інструментів існує також для редагування та модифікації ***Cases (Рядків)***, який містить операції додавання ***(Add…)***, видалення ***(Delete…)***, копіювання ***(Copy…)***, переміщення ***(Move…)***, перейменування ***(Names)*** та сортування ***(Sort)***. Ці операції можна виконати у групі ***Cases*** вкладки ***Data***.

Процес імпорту з баз даних здійснюється за допомогою стандартної опції імпорту даних або простим відкриттям файлу у системі **STATISTICA**. Така схема дозволяє уникнути проблем у разі прямого імпорту даних і похибок у процесі імпорту текстових файлів. Ці функції можливі у вкладці ***Home.***

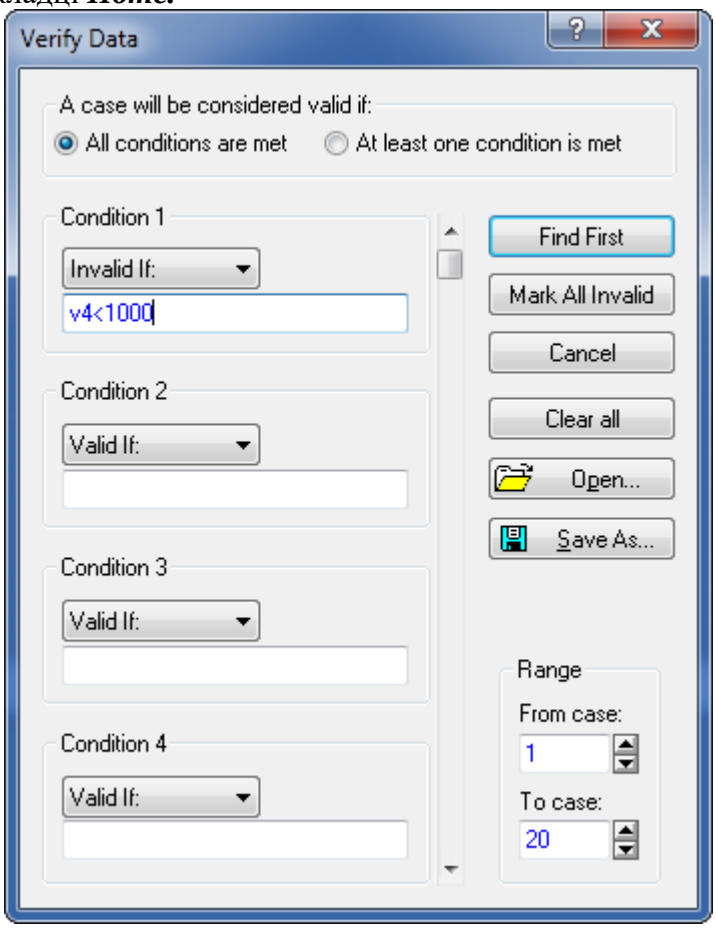


Рис. 1.12. Діалогове вікно ***Verify Data***

Система **STATISTICA** є звичайним Windows-додатком, тому дані можна зберігати, експортувавши в інший формат Windows-додатку чи конвертувавши у файл таблиць результатів.

Завдання для самостійної роботи

1.1. Запустити систему **STATISTICA**. Ознайомитися з її інтерфейсом. Вивчити призначення кожної вкладки і відповідної групи.
1.2. Створити нову електронну таблицю, яка міститиме 20 рядків і 10 стовпців.
1.3. Здійснити специфікацію електронної таблиці, ввести 20 клієнтів будь-якої організації та 10 їхніх ознак (ПІБ особи, стать, вік, освіта, середньомісячний дохід, стаж водіння, адреса тощо). Для всіх змінних (ознак) ввести назву, тип, формат відображення змінних, використовуючи кнопку ***Specs*** групи ***Variables*** вкладки ***Data***. (зокрема, середньомісячний дохід подати у грошовому форматі, адреса повинна мати довжину поля 30 тощо).
1.4. Заповнити таблицю даними.
1.5. Модифікувати створену таблицю, додавши ще будь-які 2 змінні.
1.6. Задати коди змінних стать та освіта (базова загальна середня, повна загальна середня, професійно-технічна, базова вища та повна вища освіта) за допомогою кнопки ***Text Labels*** групи ***Variables*** вкладки ***Data***.
1.7. Індексувати середньомісячний дохід (збільшити на 20%) за допомогою кнопки ***Recalculate***.
1.8. Змінну “Стаж водіння” замінити значенням рангів, використовуючи кнопку ***Rank*** групи ***Transformations*** вкладки ***Data***.
1.9. Створити нову електронну таблицю з 5 клієнтами. Злити створенні файли, використовуючи кнопку ***Merge*** групи ***Manage*** вкладки ***Data***.
1.10. Використовуючи кнопку ***Subset*** групи ***Manage*** вкладки ***Data***, відібрати клієнтів старше 45 років.
1.11. Використовуючи кнопку ***Verify Data*** групи ***Manage*** вкладки ***Data***, перевірити дані (стаж водіння < 2).
1.12. Змістити дані на два рядки і заповнити пусті, вибравши кнопку ***Shift*** групи ***Variables*** вкладки ***Data***.
1.13. Експортувати дані у MS Word або MS Excel, використовуючи кнопку ***Save As*** групи ***File*** вкладки ***Home.***

Лабораторна робота № 2
Групування статистичних даних за допомогою таблиць частот

1. Основні положення таблиць частот

Дослідження масових соціально-економічних явищ складається зі збору статистичної інформації і її первинної обробки, зведення і групування результатів спостереження у визначені сукупності, узагальнення і аналізу отриманих матеріалів. Після групування дані перетворюються у впорядковану статистичну інформацію, яка є зручною для подальшого статистичного аналізу.

Система **STATISTICA** дозволяє провести групування за різними групувальними ознаками, побудувати ряди розподілу і їх графіки. Це можливе завдяки ***Frequency tables (Таблиці частот*** (рис. 2.1) і ***Tables and banners (Таблиці і заголовки)***, що знаходяться у групі ***Base (Базові)*** вкладки ***Statistics (Статистика)*** або в меню ***Statistics (Статистика)*** модулі ***Basic Statistics and Tables (Основні базові статистики і таблиці)***.

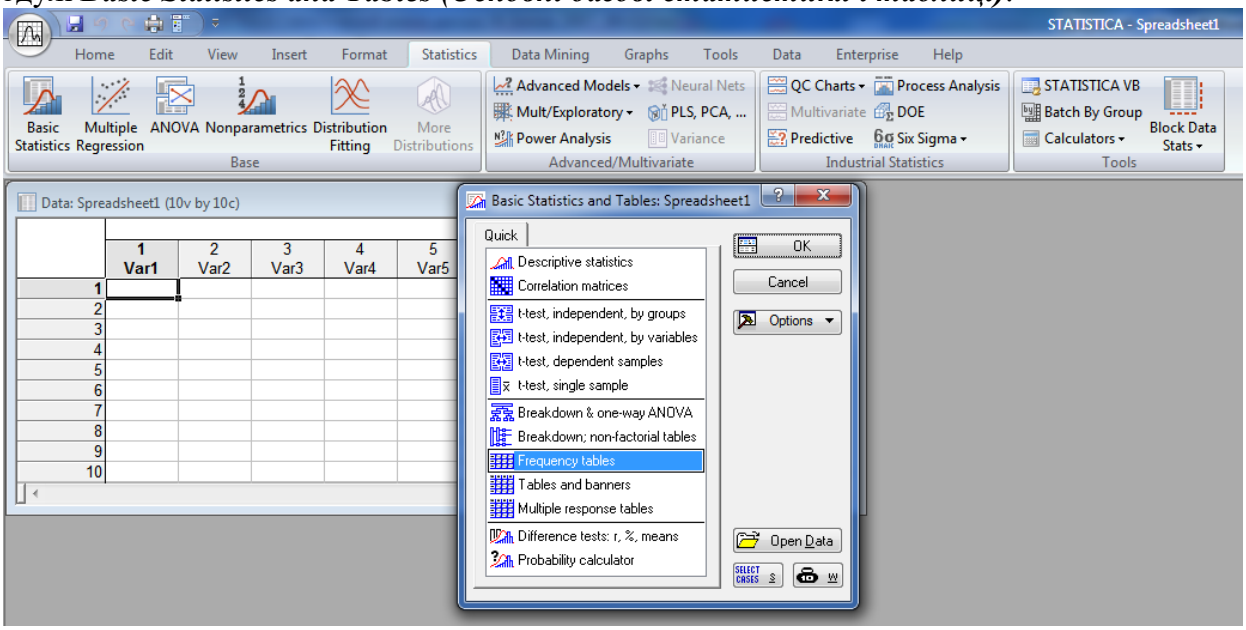


Рис. 2.1. Вікно ***Basic Statistics and Tables***

Таблиці частот є найпростішим методом статистичного аналізу, коли групування даних і побудова ряду розподілу здійснюється за однією групувальною ознакою. Діалогове вікно ***Frequency Tables*** (рис. 2.2) пропонує багато налаштувань, що дозволяють змінювати вигляд і групування в таблицях частот, а також перевіряти відповідність ряду розподілу нормальному закону, в тому числі і графічними способами.

На вкладці ***Quick (Швидкий аналіз)*** діалогового вікна ***Frequency Tables*** (рис. 2.2) розташовані такі кнопки:

- ***Variables (Змінні)*** – вибір змінних для аналізу (для вибору змінних, які розміщенні не поруч, потрібно натиснути і утримувати клавішу CTRL, клацнути мишкою на іменах потрібних змінних);
- ***Summary: Frequency tables (Результат: Таблиця частот)*** – підсумкові таблиці частот для обраних змінних;
- ***Histograms (Гістограми)*** – послідовність гістограм для обраних змінних;
- ***Descriptive Statistics (Описова статистика)*** – таблиця результатів з описовими статистиками (основними статистичними показниками) для обраних змінних;
- ***3D histograms, bivariate distributions (тривимірні гістограми, двовимірні розподіли)*** – каскад тривимірних гістограм для пар обраних змінних, один графік на кожну пару. Після вибору цієї кнопки програма попросить користувача вибрати два набори змінних (зі списку вибраних раніше за допомогою кнопки ***Variables***). Гістограми будуть побудовані для кожної пари змінних з різних списків.

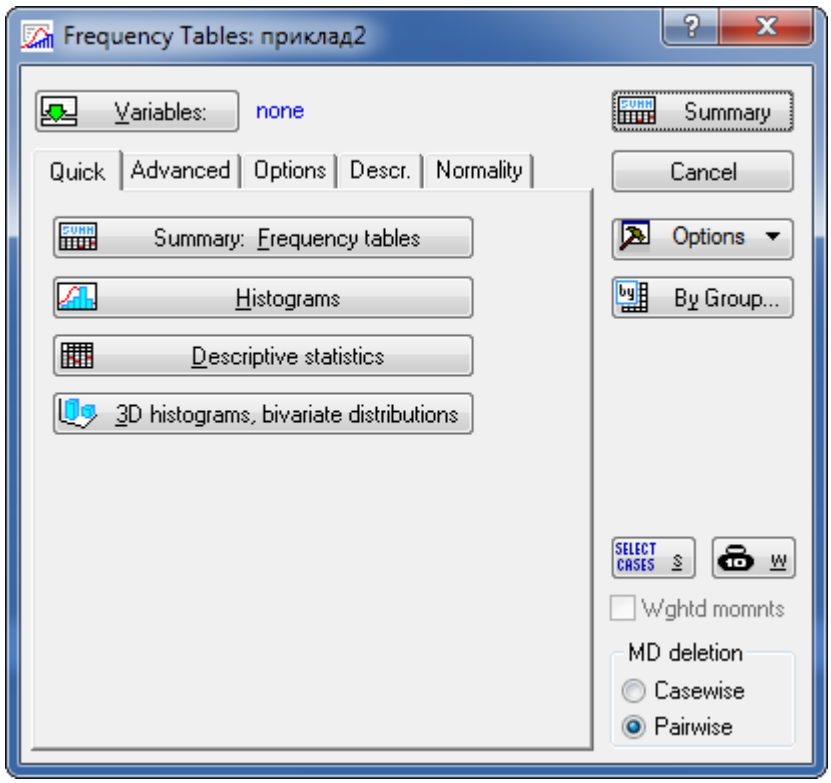


Рис. 2.2. Діалогове вікно ***Frequency Tables***

2. Додатковий аналіз таблиць частот (Advanced)

На вкладці ***Advanced (Додатковий аналіз)*** (рис. 2.3) опції під загальною назвою ***Categorization method for tables & graphs (Метод категоризації для таблиць і графіків)*** визначають, як будуть згруповані або табульовані вибрані змінні в таблицях частот і гістограмах, як обробляються спостереження при обчисленні.

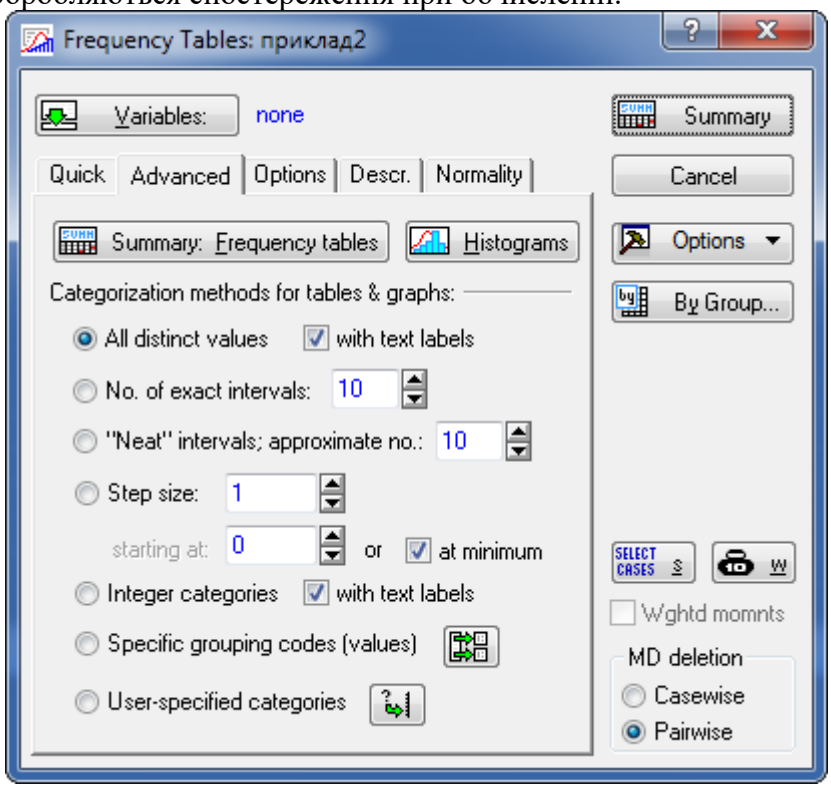


Рис. 2.3. Вкладка ***Advanced*** таблиць частот

Розглянемо всі можливості вкладки ***Advanced*** таблиць частот.

1) Опція ***All distinct values*** (***Всі різні значення)*** – дозволяє побудувати дискретний ряд розподілу. Групувальна ознака може бути як кількісною, так і якісною. Відзначивши позначкою ***with text labes (з текстовими значеннями)***, отримаємо атрибутивний ряд розподілу. В іншому випадку значення якісної змінної будуть відображені кількісно.

2) Опція ***No. of exact intervals (Кількість рівних інтервалів)*** – дозволяє побудувати ряд розподілу із заданою кількістю рівних інтервалів.

3) Опція ***«Neat» intervals; approximate no. (Наближене число інтервалів)*** – дозволяє побудувати наближені інтервали і вибере наближену довжину кроку (остання десяткова цифра буде 0 або 5, наприклад, 10,5; 11,0; 11,5 тощо).

4) Опція ***Step size (Довжина кроку)*** – дозволяє здійснити групування, спочатку задавши бажану довжину інтервалу і початок першого інтервалу, який частіше за все є мінімальним значенням ознаки (відзначити ***at minimum***) або 0 – у вікні ***starting at***, тобто почати з 0. Користувач може задати будь-яку іншу точку відліку.

5) Опція ***Integer categories (Цілі категорії)*** – будує таблицю частот тільки для цілих значень спостережень, всі нецілі значення будуть проігноровані.

У програмі передбачені і складніші способи групування даних, коли користувач сам розбиває значення ознак на класи. Наприклад, якщо встановити позначку на ***Specific grouping codes (values) (задані групуючі коди (значення))***, то таблиці частот (і гістограми) будуть побудовані цілочисельними кодами, визначеними користувачем за допомогою розташованої поряд з позначкою кнопки. Всі нецілі значення змінних будуть проігноровані програмою, а опція ***User specified categories (Визначені користувачем категорії)*** відкриває діалогове вікно, де користувач зможе здійснити свій вибір.

Графічно розподіл частот подається в системі **STATISTICA** у вигляді гістограм. Всі налаштування, які встановлені для таблиці частот, для гістограм також будуть автоматично встановлені. На гістограму також накладається гіпотетична крива нормального розподілу.

Обрати вигляд таблиці та показники, які користувач хоче бачити в ній, крім частоти, можна за допомогою групи опцій під загальною назвою ***Display options for frequency tables (Опції відображення для таблиць частот)*** розташованих на вкладці ***Options (Опції)*** діалогового вікна ***Frequency Tables***. Без вказаних додаткових опцій, таблиця частот має лише два стовпця: значення варіант і їх абсолютні частоти. Можливо обрати наступні опції відображення:

- ***Cumulative frequency (кумулятивні частоти)*** – кумулятивні (накопичені) частоти;
- ***Percentages (relative frequencies) (відсотки (відносні частоти))*** – відносні частоти (відсотки);
- ***Cumulative percentages (кумулятивні відсотки)*** – кумулятивні або накопичені відсотки;
- ***100% minus cumulative percentage (100% мінус кумулятивні відсотки)*** – значення 100 мінус кумулятивні відсотки;
- ***Logit transformed proportions (логіт-перетворення частот)*** – для частот кожної групи буде здійснено логіт-перетворення $l_i = \ln\left(\frac{n_i}{1-n_i}\right)$, де n_i – відносна частота i-ї групи;
- ***Probit transformed proportions (пробіт-перетворення частот)*** – для кумулятивних частот кожної групи буде здійснено пробіт-перетворення, завдяки якому з частот отримуємо значення, що мають нормальний розподіл;
- ***Count and report missing data (MD) (підрахунок і облік пропущених даних (ПД))*** – статистика пропущених даних;
- ***Count and report MD & non-selected cases (підрахунок і облік ПД і невибраних спостережень)*** – статистика ПД і невибраних спостережень.

Вкладка ***Descr.*** діалогового вікна ***Frequency Tables*** призначена для аналізу основних статистик досліджуваних змінних, а вкладка ***Normality*** призначена для перевірки відповідності ряду розподілу нормальному закону.

3. Типовий приклад

За даними спостережень 20 підприємств (табл. 2.1) здійснити:

а) групування підприємств за собівартістю одиниці продукції, утворивши три групи. По кожній групі визначити кількість підприємств та відсоток;

б) групування підприємств за середньою списковою чисельністю працюючих, утворивши чотири групи з наближеними інтервалами. По кожній групі знайти абсолютні частоти, відносні частоти, кумулятивні абсолютні частоти та кумулятивні відносні частоти.

в) групування підприємств за вартістю основних виробничих засобів і виробництвом продукції, утворивши групи з інтервалами з довжиною три та п'ять відповідно. У кожній групі обчислити кількість підприємств.

Результати групувань подати у вигляді таблиць та гістограм.

Таблиця 2.1

№ з/п	Середньорічна вартість основних виробничих засобів (млн. грн.)	Вироблено продукції за звітний місяць (млн. грн.)	Собівартість одиниці продукції (грн.)	Середня списокова чисельність працюючих (осіб)
1	26,1	33,3	71	360
2	21,6	29,8	75	275
3	22,0	21,6	72	313
4	35,0	44,4	67	475
5	13,4	16,5	86	200
6	22,9	27,9	66	286
7	16,7	19,6	76	291
8	26,8	22,0	65	418
9	32,1	36,7	64	375
10	24,8	23,6	71	340
11	12,9	14,5	92	109
12	11,4	16,6	95	257
13	18,0	21,5	93	215
14	33,2	43,0	60	421
15	23,1	32,6	73	240
16	11,5	18,0	86	120
17	9,2	13,7	89	115
18	15,7	28,5	74	252
19	13,6	12,5	94	280
20	31,4	42,9	61	410

Розв'язування. *Для здійснення групування підприємств за собівартістю одиниці продукції, потрібно вибрати змінну* ***Собівартість одиниці продукції (грн.)*** *у діалоговому вікні* ***Frequency Tables*** *(рис. 2.4), а потім у вкладці* ***Advanced*** *задати* ***Кількість рівних інтервалів (No. of exact intervals)*** *– 3 (рис. 2.5). Результат такого групування відображено на рис. 2.6.*

Для побудови гістограми потрібно вибрати кнопку ***Histograms*** *у діалоговому вікні* ***Frequency Tables****. Гістограма групування зображена на рис. 2.7.*

Для здійснення групування підприємств за середньою списковою чисельністю працюючих потрібно вибрати змінну ***Середня списокова чисельність працюючих (осіб)*** *у діалоговому вікні* ***Frequency Tables*** *(рис. 2.4), а потім у вкладці* ***Advanced*** *задати* ***«Neat» intervals; approximate no. (Наближене число інтервалів)*** *– 4 (рис. 2.8) та на вкладці* ***Options*** *вибрати* ***Cumulative frequency, Percentages (relative frequencies) та Cumulative percentages*** *(рис. 2.9) для відображення відносних частот, кумулятивних абсолютних частот та кумулятивних відносних частот в таблиці результатів (рис. 2.10).*

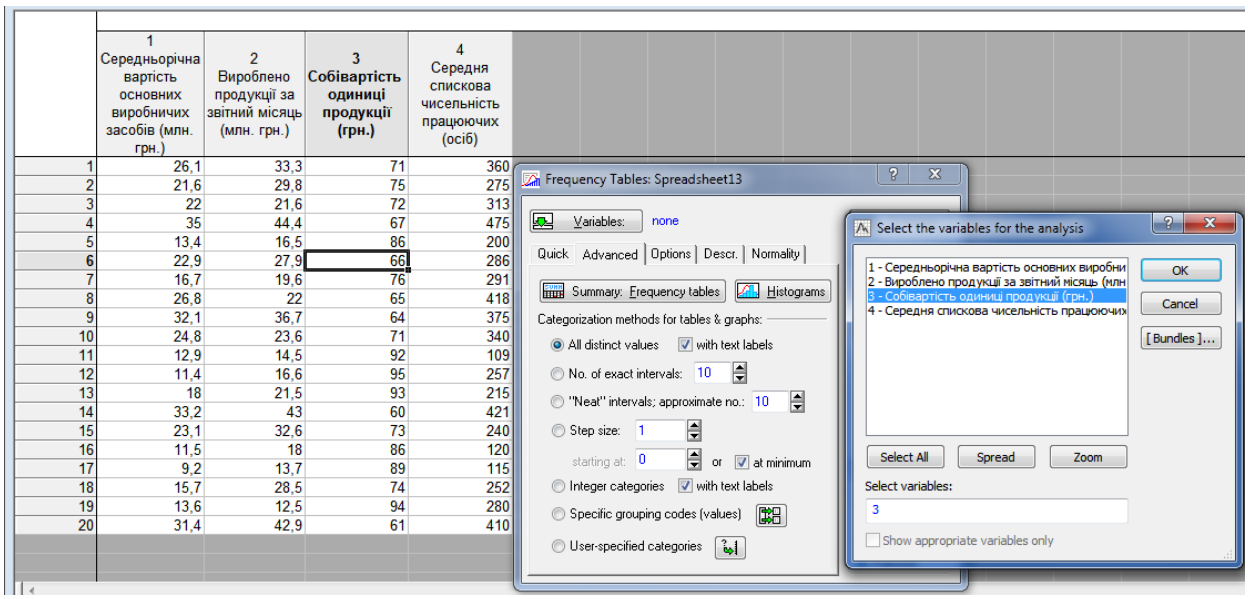


Рис. 2.4. Вибір змінної для групування

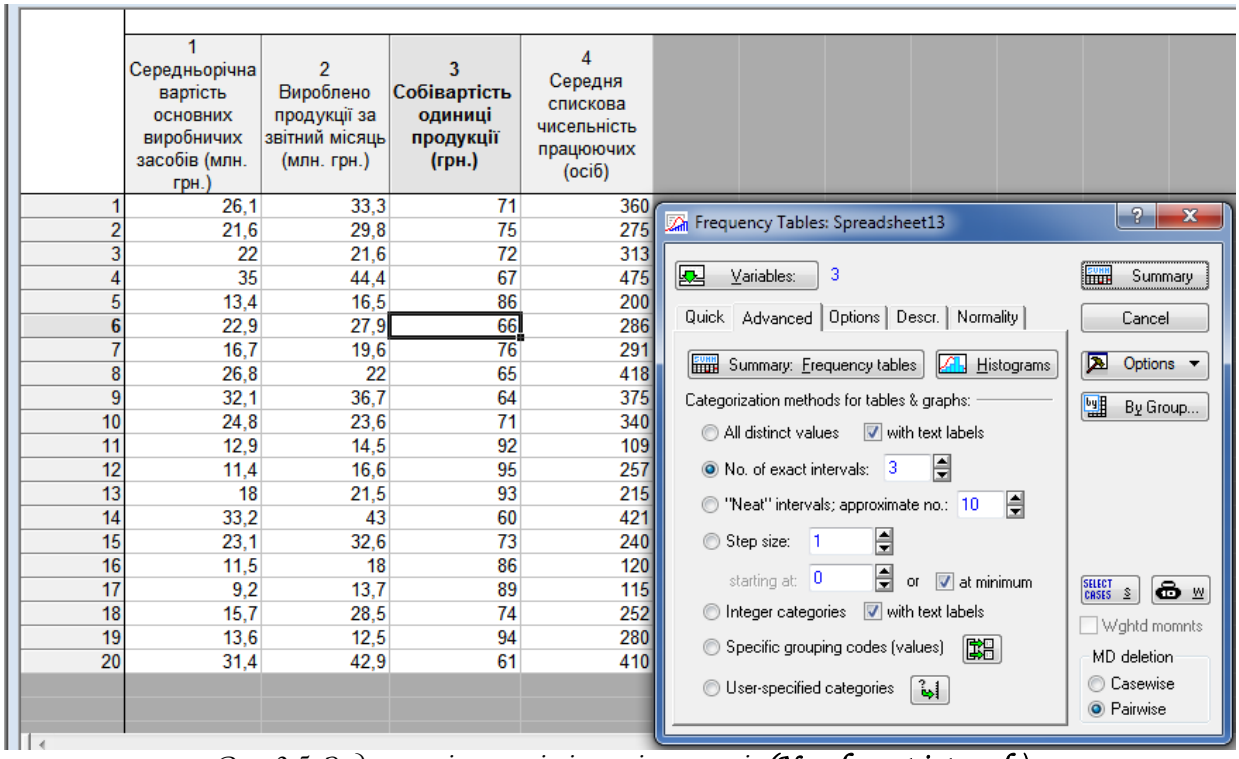


Рис. 2.5. Задання кількості рівних інтервалів ***(No. of exact intervals)***

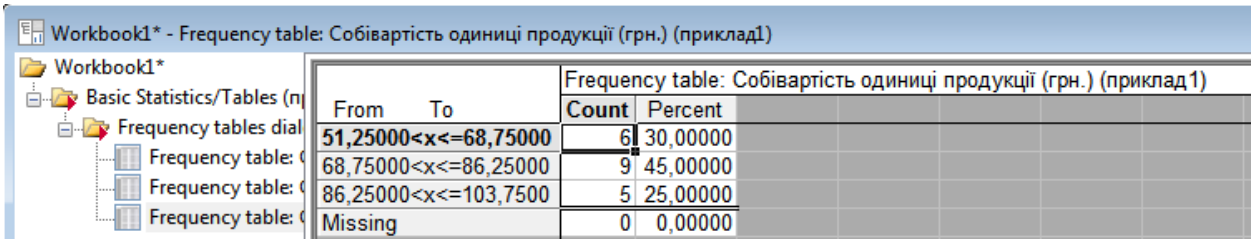


Рис. 2.6. Результат групування за собівартістю одиниці продукції

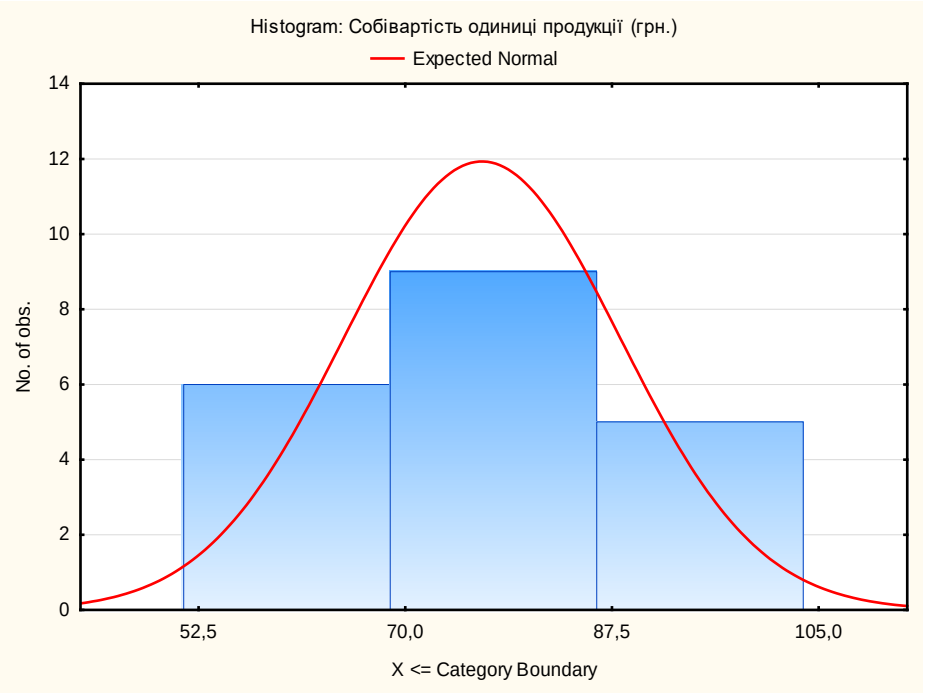


Рис. 2.7. Гістограма групування за собівартістю одиниці продукції

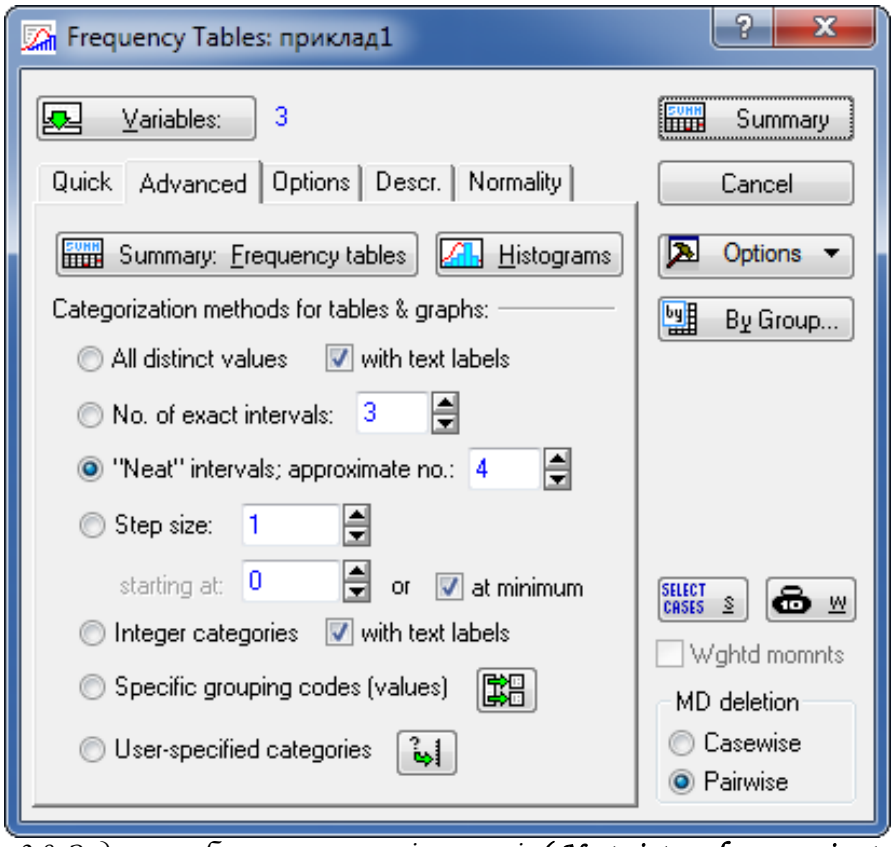


Рис. 2.8. Задання наближеного числа інтервалів ***(«Neat» intervals; approximate no.)***

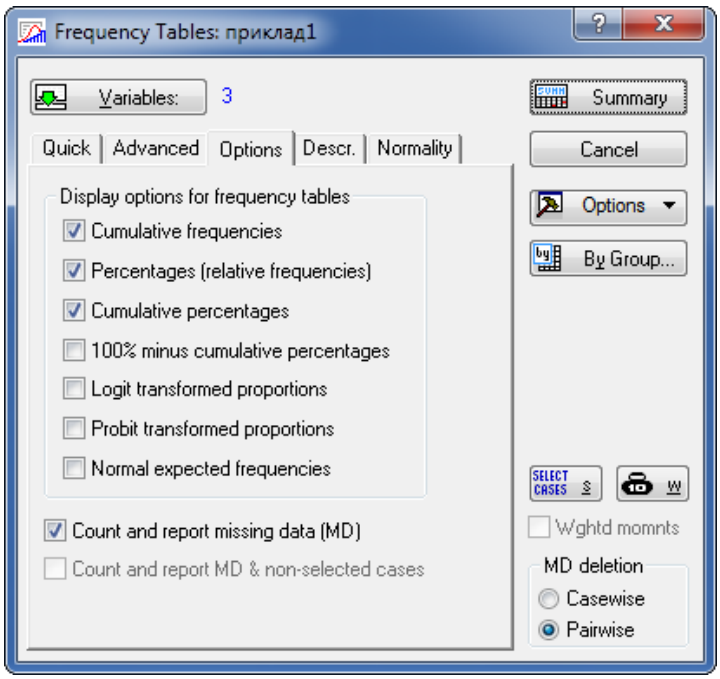


Рис. 2.9. Задання показників групування даних

	Frequency table: Собівартість одиниці проду			
From To	Count	Cumulative Count	Percent	Cumulative Percent
50,00000<x<=60,00000	1	1	5,00000	5,0000
60,00000<x<=70,00000	5	6	25,00000	30,0000
70,00000<x<=80,00000	7	13	35,00000	65,0000
80,00000<x<=90,00000	3	16	15,00000	80,0000
90,00000<x<=100,0000	4	20	20,00000	100,0000
100,0000<x<=110,0000	0	20	0,00000	100,0000
Missing	0	20	0,00000	100,0000

Рис. 2.10. Результат групування за середньою списковою чисельністю працюючих

Аналогічно задаються інші умови групування (рис. 2.11). ***Увага****: Результати роботи з'являються у окремому файлі –* ***Workbook (Робочій книзі)*** *(розширення файлу – *.stw). У системі* ***Statistica*** *таблиця вхідних даних та результати роботи зберігаються окремими файлами з відповідними розширеннями.*

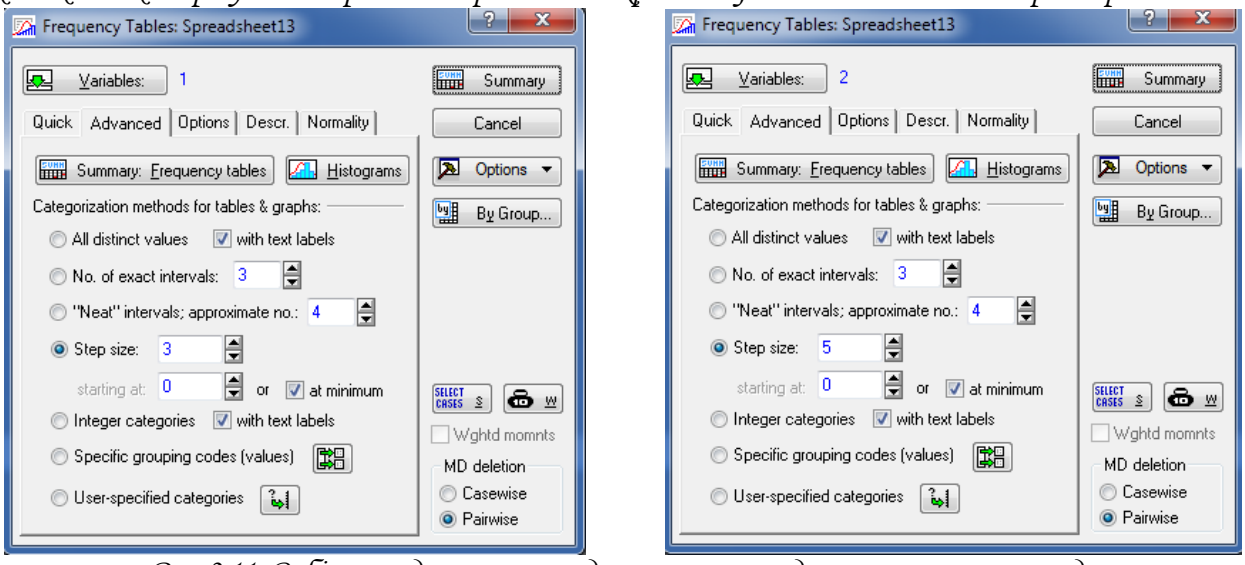


Рис. 2.11. Вибір методу групування для третього завдання типового прикладу

Завдання для самостійної роботи

2.1. На основі вибіркової сукупності даних (табл. 2.2) про одночасні покупки соків і фруктів в супермаркеті 50 покупців чоловічої і жіночої статі визначити:

1) відсоткове співвідношення чоловіків і жінок, що взяли участь в опитуванні; частку (в %) кожного виду соків (фруктів) в загальному обсязі куплених соків (фруктів);

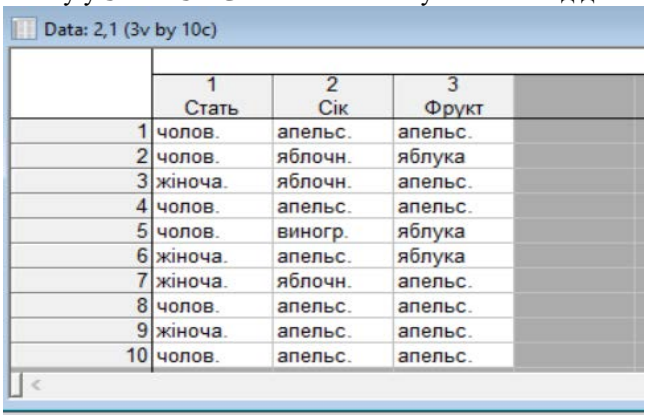
2) частоти переваги різних фруктів на кожному рівні переваги соків: а) по всіх респондентах в цілому; б) окремо по чоловіках і жінках. (***Вказівка:*** *необхідно створити новий файл зі змінною – чоловіки і провести аналіз, потім слід створити інший файл – змінна – жінки і провести аналогічний аналіз).* Порівняти результати аналізу по пунктах а) і б).

Таблиця 2.2

Стать	*Сік*	*Фрукт*
чолов.	апельс.	апельс.
чолов.	яблочн.	яблука
жіноча.	яблочн.	апельс.
чолов.	апельс.	апельс.
чолов.	виногр.	яблука
жіноча.	апельс.	яблука
жіноча.	яблочн.	апельс.
чолов.	апельс.	апельс.
жіноча.	апельс.	апельс.
чолов.	апельс.	апельс.
жіноча.	виногр.	виногр.
чолов.	виногр.	виногр.
чолов.	апельс.	яблука
чолов.	яблочн.	апельс.
жіноча.	апельс.	виногр.
жіноча.	апельс.	виногр.
чолов.	яблочн.	апельс.
жіноча.	виногр.	виногр.
жіноча.	яблочн.	апельс.
чолов.	апельс.	яблука
чолов.	апельс.	апельс.
жіноча.	виногр.	виногр.
чолов.	виногр.	виногр.
жіноча.	апельс.	апельс.
чолов.	яблочн.	яблука

Стать	*Сік*	*Фрукт*
чолов.	апельс.	виногр.
чолов.	апельс.	яблука
жіноча.	апельс.	апельс.
жіноча.	апельс.	апельс.
чолов.	виногр.	яблука
чолов.	апельс.	апельс.
жіноча.	яблочн.	яблука
жіноча.	апельс.	апельс.
чолов.	яблочн.	яблука
чолов.	виногр.	виногр.
чолов.	виногр.	апельс.
жіноча.	виногр.	апельс.
жіноча.	апельс.	яблука
жіноча.	апельс.	яблука
чолов.	яблочн.	апельс.
чолов.	апельс.	яблука
чолов.	яблочн.	яблука
жіноча.	апельс.	апельс.
чолов.	апельс.	апельс.
чолов.	яблочн.	яблука
чолов.	виногр.	виногр.
жіноча.	виногр.	виногр.
жіноча.	апельс.	яблука
жіноча.	яблочн.	апельс.
чолов.	яблочн.	яблука

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх 50 спостережень.


Data: 2,1 (3v by 10c)

	1 Стать	2 Сік	3 Фрукт
1	чолов.	апельс.	апельс.
2	чолов.	яблочн.	яблука
3	жіноча.	яблочн.	апельс.
4	чолов.	апельс.	апельс.
5	чолов.	виногр.	яблука
6	жіноча.	апельс.	яблука
7	жіноча.	яблочн.	апельс.
8	чолов.	апельс.	апельс.
9	жіноча.	апельс.	апельс.
10	чолов.	апельс.	апельс.

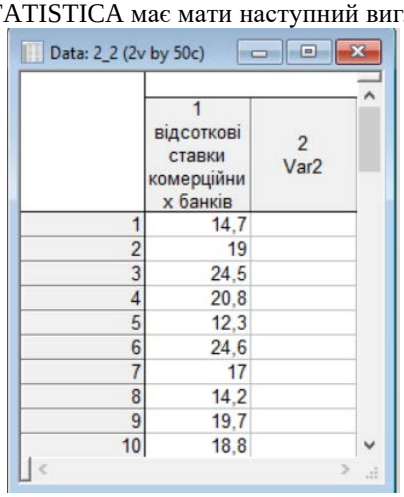
2.2. У результаті статистичного спостереження комерційних банків отримано дані (табл. 2.3) про відсоткові ставки комерційних банків (%).

Таблиця 2.3

14,7	19,0	24,5	20,8	12,3	24,6	17,0	14,2	19,7	18,8
18,1	20,5	21,0	20,7	20,4	14,7	25,1	22,7	19,0	19,6
19,0	18,9	17,4	20,0	13,8	25,6	13,0	19,0	18,7	21,1
13,3	20,7	15,2	19,9	21,9	16,0	16,9	15,3	21,4	20,4
12,8	20,0	14,3	18,0	15,1	23,8	18,5	14,4	21,0	19,0

Здійснити групування комерційних банків на основі відсоткової ставки, задавши п’ять рівних інтервалів. Який інтервал має найбільшу абсолютну частоту? Перегрупувати дані, використовуючи опцію ***«Neat» intervals app.no***. Порівняти обидва результати здійснених групувань. Вибрати кращий і пояснити.

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх 50 спостережень.



2.3. За даними про підприємства ресторанного господарства (позначено: Р – ресторани; К – кафе; Б – бари; Ї – їдальні) (табл. 2.4) побудувати ряд розподілу підприємств за їх видами та подати його графічно. Проаналізувати отримані результати.

Таблиця 2.4

К	К	Б	К	К	К	К	К	Б	Ї
Б	Р	К	Ї	К	Ї	К	Ї	К	К
Б	Б	К	К	К	К	Р	Б	Ї	К
К	К	Б	К	К	Р	К	К	К	К
Ї	К	К	Б	Б	К	Р	Ї	Р	Ї
Р	К	К	Р	Ї	Б	К	К	К	К
К	К	Ї	К	К	К	К	К	К	Б

Примітка. Файл для аналізу у STATISTICA має бути сформований для 1 змінної та 70 спостережень (аналогічно до 2.2).

2.4. Результати семестрової роботи з навчальної дисципліни “Статистика” чотирьох груп (за 100-бальною шкалою) наведені в таблиці 2.5.

Таблиця 2.5

Гр-1	Гр-2	Гр-3	Гр-4
78; 45; 94; 47; 50; 57; 49; 90; 67; 67; 76; 52; 49; 60; 59; 90; 95; 65; 68; 64; 98; 68; 64	91; 60; 96; 63; 81; 72; 36; 79; 48; 50; 36; 81; 58; 58; 66; 54; 56; 50; 56; 63; 62; 57; 100; 80; 54; 83; 72; 85; 74	100; 45; 51; 89; 94; 62; 50; 70; 71; 67; 79; 64; 62; 71; 80; 64; 92; 44; 71; 53; 65; 100; 44; 93; 94; 100; 100	100; 95; 100; 98; 53; 66; 67; 52; 83; 86; 64; 48; 77; 72; 49; 58; 49; 72; 68; 63; 100; 66

За даними обстеження: 1) згрупувати студентів за набраними балами у ряд розподілу з рівними інтервалами; 2) обчислити відносні частоти; 3) згрупувати студентів за оцінками (за

шкалою ECTS); 4) подати утворені ряди розподілу графічно та таблично. Проаналізувати отримані результати.

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх спостережень.

Data: 2_4 (4v by 10c)

	1 Гр-1	2 Гр-2	3 Гр-3	4 Гр-4
1	78	91	100	100
2	45	60	45	95
3	94	96	51	100
4	47	63	89	98
5	50	81	94	53
6	57	72	62	66
7	49	36	50	67
8	90	79	70	52
9	67	48	71	83
10	67	50	67	86

2.5. У таблиці 2.6 подано дані про характеристики 28 автомобілів.

Таблиця 2.6

	Об’єм двигуна (л)	Потужність двигуна (к.с.)	Роздрібна ціна (тис. EUR)
VW Passat	2,0	115	22,3
Mersedes C200 kompressor	1,8	163	32,1
Saab 9-3 Sport Combi	2,0	175	25,0
Mazda 6	2,0	143	21,6
Honda Jazz	1,4	83	16,1
Scoda Fabia	1,4	75	16,0
BMW 3 Touring	2,0	143	32,0
Audi A4 Avant	1,8	163	32,5
Chevrolet Evanda	2,0	131	13,1
BMW 318 i	2,0	143	27,3
Citroën Berlingo	2,0	90	11,8
KIA Sorrento	2,4	139	30,9
Mersedes C203	1,8	163	38,5
Mazda Premacy	2,0	131	22,5
Peugeot 407	2,0	138	25,0
Hyundai Terracan	2,9	150	34,1
Volvo XC90 V8	4,4	315	82,0
Peugeot Partner	1,9	70	12,1
Mitsubishi Colt	1,3	95	12,5
Volvo V50 N5 AWD	2,5	220	42,0
Mitsubishi Pajero Sport	3,0	170	32,5
Ford Connect	1,8	90	16,5
BMW X5 4.4	4,4	286	70,0
Honda Accord	2,4	196	26,5
Ford Mondeo	2,0	145	25,0
Audi A8	3,0	218	68,1
Peugeot 206	1,1	60	10,0
Toyota Avensis	1,8	129	25,7

Здійснити групування: а) за об’ємом двигуна методом рівних інтервалів; б) за роздрібною ціною, утворивши 5 рівних інтервалів; в) за потужністю двигуна з кроком 5. Побудувати таблиці частот та проаналізувати отримані результати.

Лабораторна робота № 3
Групування статистичних даних за допомогою кростабуляції

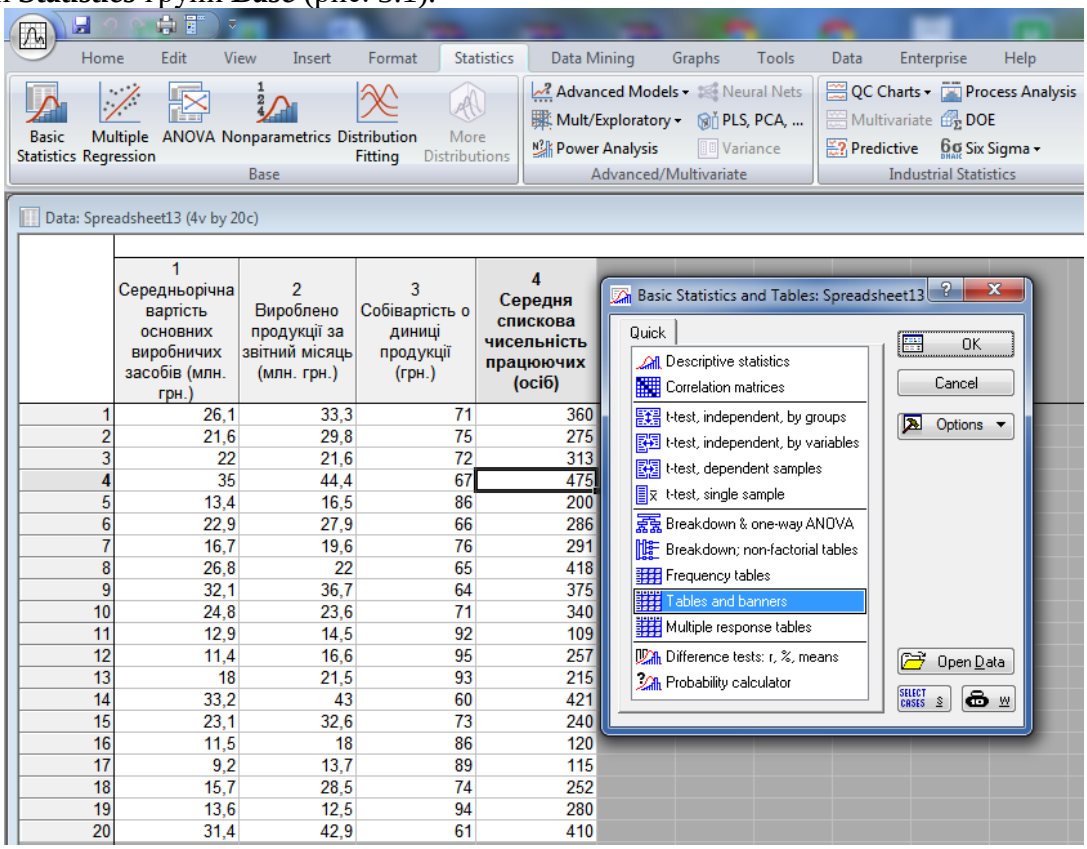
1. Основні положення кростабуляції

Кростабуляція – це статистичний метод, при якому одночасно аналізуються значення двох чи більше змінних. Кростабуляція полягає у створенні таблиць взаємної спряженості ознак, які відображають сумісний розподіл двох чи більше змінних з обмеженою кількістю категорій або визначеними значеннями.

Використання кростабуляції дозволяє, наприклад, вирішити проблему залежності між відповідями на різні питання анкет при соціологічному чи маркетинговому дослідженні тощо.

Звичайно табулюються категоріальні змінні або змінні з обмеженою кількістю значень. Якщо необхідно табулювати неперервну змінну (наприклад, дохід), то спочатку її слід перекодувати, розбивши діапазон зміни на невелике число інтервалів (наприклад, дохід: низький, середній, високий). Слід пам'ятати, що кростабуляція працює тільки з цілоцисельними значеннями.

Crosstabulation tables (Таблиці кростабуляції) будуються за допомогою вибору ***Tables and banners (Таблиці і заголовки)*** зі стартової панелі модуля ***Basic Statistics and Tables*** вкладки ***Statistics*** групи ***Base*** (рис. 3.1).



	1 Середньорічна вартість основних виробничих засобів (млн. грн.)	2 Вироблено продукції за звітний місяць (млн. грн.)	3 Собівартість одиниці продукції (грн.)	4 Середня списокова чисельність працюючих (осіб)
1	26,1	33,3	71	360
2	21,6	29,8	75	275
3	22	21,6	72	313
4	35	44,4	67	475
5	13,4	16,5	86	200
6	22,9	27,9	66	286
7	16,7	19,6	76	291
8	26,8	22	65	418
9	32,1	36,7	64	375
10	24,8	23,6	71	340
11	12,9	14,5	92	109
12	11,4	16,6	95	257
13	18	21,5	93	215
14	33,2	43	60	421
15	23,1	32,6	73	240
16	11,5	18	86	120
17	9,2	13,7	89	115
18	15,7	28,5	74	252
19	13,6	12,5	94	280
20	31,4	42,9	61	410

Рис. 3.1. Вибір модуля кростабуляції

Відкриється діалогове вікно ***Crosstabulation Tables (Таблиці кросстабуляції)***, що містить дві вкладки ***Stub-and-banner table (Таблиці прапорів і заголовків)*** – побудова таблиці взаємоспряженості для двох змінних і ***Crosstabulation (Кростабуляція)*** – побудова таблиць взаємоспряженості для більше, ніж дві змінні (рис. 3.2).

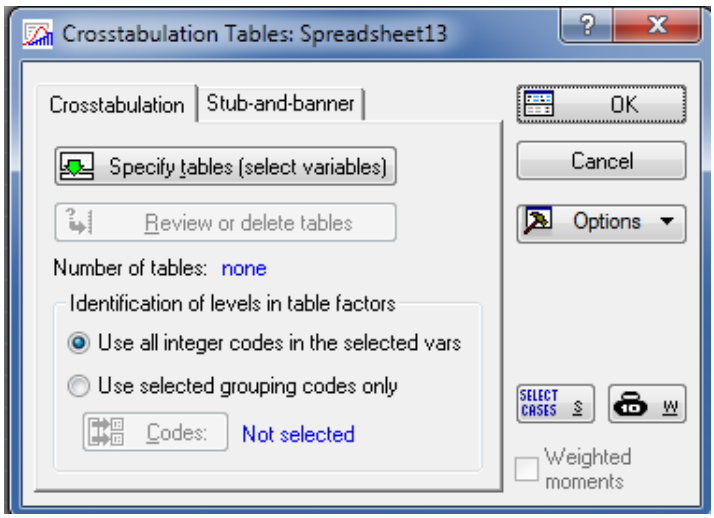


Рис. 3.2. Діалогове вікно ***Crosstabulation Tables***

На вкладці ***Crosstabulation*** треба натиснути кнопку ***Specify tables (Специфікувати таблицю)*** та вибрати змінні для аналізу. Найпростіший випадок кростабуляції – дослідження двох змінних, що мають категоріальні або цілочислові значення. Однак, **STATISTICA** дозволяє аналізувати одночасно до шести пар змінних (рис. 3.3).

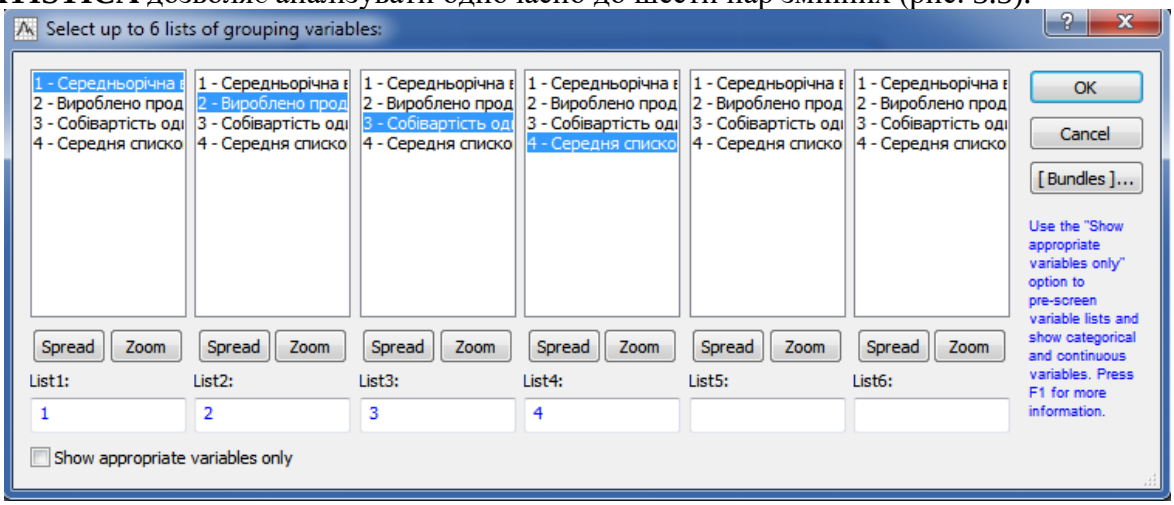


Рис. 3.3. Меню вибору змінних для кростабуляції

Після вибору змінних для побудови таблиць у полі ***Number of tables (Кількість таблиць)*** діалогового вікна ***Crosstabulation Tables*** буде показано кількість таблиць, в яких будуть відображені результати кростабуляції. Можна переглянути їх назви та вибрати для подальшого аналізу лише необхідні, скориставшись опцією ***Review or delete tables (Переглянути чи видалити таблиці)*** (рис. 3.4).

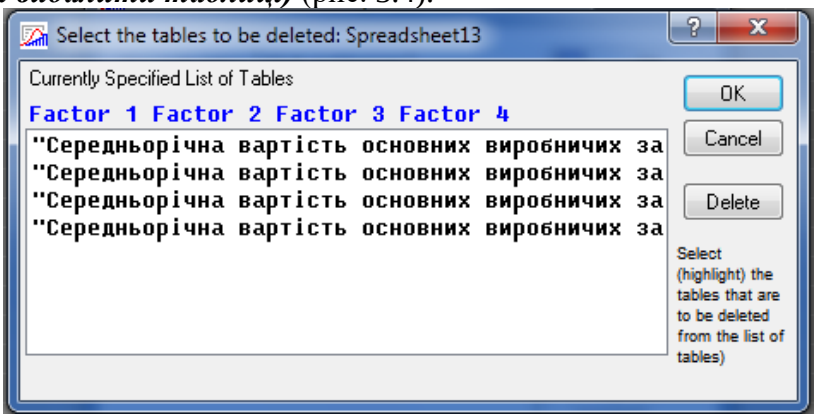


Рис. 3.4. Меню виробу таблиць для аналізу

Необхідно також вказати категорії, які будуть використовуватися. Якщо використовуватимуться всі категорії в обраних змінних, потрібно позначити опцію ***Use all integer codes in the selected vars (Використовувати всі цілі значення (коди) в обраних змінних)***, якщо деякі – ***Use selected grouping codes only (Використовувати тільки обрані коди)***. У полі ***Codes (Коди)*** позначають ті категорії, які необхідні для аналізу. Кнопкою ***Zoom*** викликається перегляд усіх категорій у змінній.

Вкладка ***Stub-and-banner tables (Таблиці прапорів і заголовків)*** діалогового вікна ***Crosstabulation Tables*** дозволяє відобразити таблицю, яка побудована тільки для двох змінних.

2. Методи подання результатів кростабуляції

Після вибору всіх параметрів крос табуляції натискаємо ***OK***, після чого з'явиться меню виводу результатів кростабуляції (рис. 3.5).

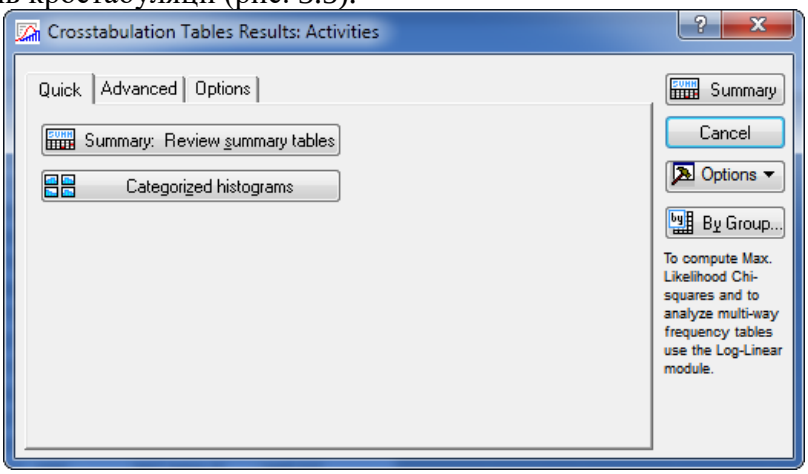


Рис. 3.5. Меню виводу результатів кростабуляції (вкладка ***Quick***)

У меню виводу результатів можна вибрати наступні види подання результатів:

- ***Quick (Швидкий аналіз)***;
- ***Advanced (Додатковий аналіз)***;
- ***Options (Опції)***.

Вкладка ***Quick*** діалогового вікна ***Crosstabulation Tables Results*** містить такі кнопки:

- ***Summary: Review summary tables (Переглянути підсумкові таблиці)*** – побудова підсумкових таблиць. Якщо побудовано більше однієї таблиці, то відкривається діалогове вікно, в якому можна вибрати таблиці для перегляду. Якщо вибрано ***All tables (Всі таблиці)***, буде побудований каскад таблиць результатів кростабуляції.
- ***Categorized histograms (Категоризовані гістограми)*** – побудова гістограм. Якщо побудовано більше, ніж одна таблиця, то можна вибрати, значення з якої саме таблиці відобразити на гістограмі. Зазначимо, що кожна гістограма може містити інформацію максимум про три змінні (фактори).

Вкладка ***Advanced*** діалогового вікна ***Crosstabulation Tables Results*** (рис. 3.6) дозволяє вибрати:

- ***Summary: Review summary tables***.
- ***Detailed Two-way Tables (Докладні таблиці для двох змінних)*** – таблиці результатів для двох змінних.
- ***Categorized histograms.***
- ***Interaction plots of frequencies (Графіки взаємодії частот)*** – лінійний графік (графік взаємодії), який показує розподіл частот між змінними (факторами). Якщо в таблиці є більше трьох факторів, то програма побудує декілька графіків.
- ***3D histograms*** (***Тривимірні гістограми***) – 3D гістограми для обраних таблиць.
- ***Display Long Text Labels (Відображати довгі текстові значення)*** – відображення довгих текстових значень у першому стовпці таблиці для двох змінних. Якщо

відповідний фактор не має довгої назви, опція ігнорується.

- ***Include Missing Data (Враховувати пропущені значення)***.
- ***Display Selected %'s in Sep. Tables (Відображати вибрані % в окремих таблицях)*** – процедура доступна, якщо вибрана одна з опцій ***Percentages… (Відсотки…)*** в полі ***Compute tables (Підрахувати таблиці)*** вкладки ***Options*** діалогового вікна ***Crosstabulation Tables Results***. За замовчуванням, відсотки будуть відображені в одній таблиці з частотами. Якщо виділена дана опція, то таблиці з відсотками будуть виведені на екран окремо.

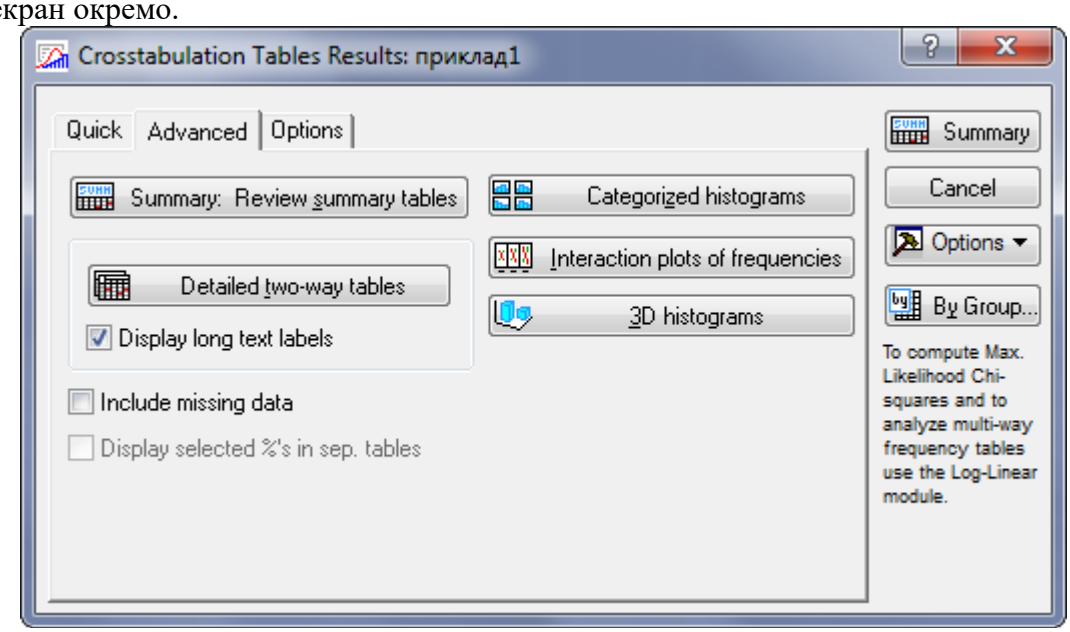


Рис. 3.6. Вкладка ***Advanced*** діалогового вікна ***Crosstabulation Tables Results***

Вкладка ***Options*** діалогового вікна ***Crosstabulation Tables Results*** дозволяє вибрати (рис. 3.7):

- ***Compute tables (Підрахувати таблиці)*** – можливість виведення результатів в ***докладних таблицях для двох змінних***, ***кростабуляційних таблицях*** і в ***таблицях прапорів і заголовків***;
- ***Statistics for two-way tables (Статистики для таблиць, які побудовані для двох змінних)*** – основні статистики для таблиць з двома змінними.

У свою чергу ***Compute tables*** містить поля, в яких за необхідністю ставлять позначки:

- ✓ ***Highlight counts> (Виділити частоти>)*** – всі частоти по рядках, які перевищують введене значення, будуть виділені червоним кольором;
- ✓ ***Expected frequencies (Очікувані частоти)*** – очікувані частоти в припущенні про незалежність всіх факторів (змінних) в таблиці з двома змінними;
- ✓ ***Residual frequencies (Залишкові частоти)*** – залишкові частоти (спостережувані частоти мінус очікувані частоти) для всіх типів таблиць;
- ✓ ***Percentages of total count (Відсотки від загального числа)*** – відсотки для кожної комірки;
- ✓ ***Percentages of row counts (Відсотки по рядку)*** – відсотки відносно загальної кількості спостережень у відповідному рядку поточної таблиці для кожної комірки;
- ✓ ***Percentages of column counts (відсотки по стовпцю)*** – аналогічно до попереднього, тільки стосовно стовпців таблиці.

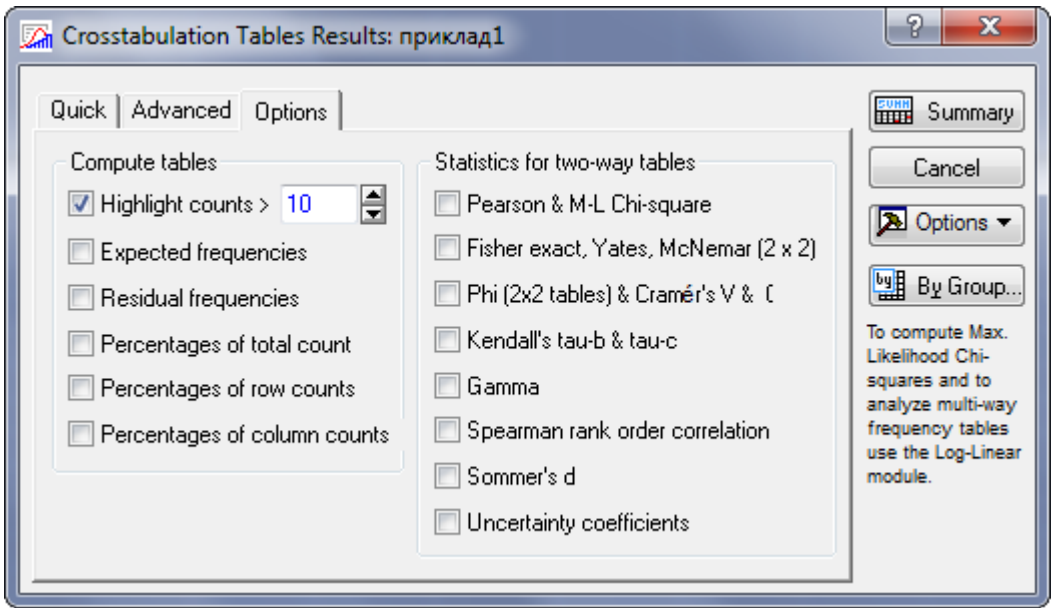


Рис. 3.7. Вкладка ***Options*** діалогового вікна ***Crosstabulation Tables Results***

Statistics for two-way tables дозволяє додатково знайти основні статистики таблиць, які побудовані для двох змінних:

- ✓ ***Pearson Chi-square*** – критерій Пірсона (χ^2 -критерій);
- ✓ ***M-L Chi-square*** – критерій максимальної правдоподібності;
- ✓ ***Yates correction*** – критерій згоди Пірсона з поправкою Йєтса на неперервність;
- ✓ ***Fisher exact test*** – точний тест Фішера;
- ✓ ***McNemar Chi-square*** – критерій Макнемара (застосовується, коли частоти в таблиці 2x2 подані залежні вибірки);
- ✓ ***Phi (2x2 tables) & Cramer's V&C*** – ϕ -критерій;
- ✓ ***Kendall tau*** – коефіцієнт рангової кореляції Кендала;
- ✓ ***Gamma*** – гамма статистика (якщо дані мають багато однакових значень, ця статистика краще, ніж коефіцієнт Спірмена або Кендала);
- ✓ ***Spearman rank order correlation*** – коефіцієнт рангової кореляції Спірмена;
- ✓ ***Sommer's d*** – статистика Соммера (несиметрична міра зв'язку між двома змінними);
- ✓ ***Uncertainty Coefficients*** – коефіцієнти невизначеності (вимірюють інформаційний зв'язок між факторами (рядками і стовпцями таблиці)).

Вказані статистики будуть подані в ***таблицях для двох змінних та кростабуляційних таблицях*** (рис. 3.8).

	Statistics: Y_1983(9) x NewVar1(8) (Ac		
Statistic	**Chi-square**	df	p
Pearson Chi-square	59,00000	df=56	p=,36643
M-L Chi-square	39,18314	df=56	p=,95717
Phi	2,217356		
Contingency coefficient	,9115843		
Cramer's V	,8380817		
Kendall's tau b & c	b=,7741935	c=,7619048	
Sommers D(X|Y), D(Y|X)	X|Y=,77419	Y|X=,77419	
Gamma	,8275862		
Spearman Rank R	,8863246	t=6,0527	p=,00012
Uncertainty coefficient	X=,7793996	Y=,8071110	X|Y=,79301

Y_1983: Number of accidents in 1983	NewVar1 40	NewVar1 49	NewVar1 50	NewVar1 60	NewVar1 74	NewVar1 80	NewVar1 89	NewVar1 100	Row Totals
40	1	0	0	0	0	0	0	0	1
Column Percent	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Row Percent	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Total Percent	8,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	8,33%
43	0	0	1	0	0	0	0	0	1
Column Percent	0,00%	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Row Percent	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Total Percent	0,00%	0,00%	8,33%	0,00%	0,00%	0,00%	0,00%	0,00%	8,33%
50	1	1	0	0	0	0	0	0	2
Column Percent	50,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Row Percent	50,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Total Percent	8,33%	8,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	16,67%
65	0	0	0	1	0	0	0	0	1
Column Percent	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	
Row Percent	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	
Total Percent	0,00%	0,00%	0,00%	8,33%	0,00%	0,00%	0,00%	0,00%	8,33%
75	0	0	0	0	1	0	0	0	1
Column Percent	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	
Row Percent	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	
Total Percent	0,00%	0,00%	0,00%	0,00%	8,33%	0,00%	0,00%	0,00%	8,33%
80	0	0	1	0	0	1	1	0	3
Column Percent	0,00%	0,00%	50,00%	0,00%	0,00%	50,00%	100,00%	0,00%	
Row Percent	0,00%	0,00%	33,33%	0,00%	0,00%	33,33%	33,33%	0,00%	
Total Percent	0,00%	0,00%	8,33%	0,00%	0,00%	8,33%	8,33%	0,00%	25,00%
95	0	0	0	0	0	1	0	0	1
Column Percent	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	0,00%	0,00%	
Row Percent	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	
Total Percent	0,00%	0,00%	0,00%	0,00%	0,00%	8,33%	0,00%	0,00%	8,33%
125	0	0	0	0	0	0	0	1	1
Column Percent	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	
Row Percent	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	
Total Percent	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	8,33%	8,33%
150	0	0	0	0	0	0	0	1	1
Column Percent	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	
Row Percent	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	
Total Percent	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	8,33%	8,33%
Totals	2	1	2	1	1	2	1	2	12
Total Percent	16,67%	8,33%	16,67%	8,33%	8,33%	16,67%	8,33%	16,67%	100,00%

2-Way Summary Table: Observed Frequencies (Accident2)
Marked cells have counts > 10

Рис. 3.8. Результати аналізу і групування кростабуляції

3. Типовий приклад

За даними про якість ґрунтів та урожайність озимої пшениці по тридцяти сільськогосподарських підприємствах регіону (таблиця 3.1) виконати аналітичне групування для вивчення залежності урожайності озимої пшениці від якості ґрунтів за допомогою кростабуляції. Результати кростабуляції подати у вигляді таблиці та графічно. Проаналізувати результати групування.

Таблиця 3.1

№ з/п	*Середня якість ґрунту (балів)*	*Урожайність озимої пшениці, ц/га*	*№ з/п*	*Середня якість ґрунту (балів)*	*Урожайність озимої пшениці, ц/га*
1	*41*	*29*	*16*	*41*	*27*
2	*46*	*36*	*17*	*46*	*32*
3	*50*	*39*	*18*	*54*	*41*
4	*53*	*40*	*19*	*43*	*30*
5	*42*	*25*	*20*	*47*	*33*
6	*48*	*37*	*21*	*49*	*34*
7	*45*	*32*	*22*	*52*	*41*
8	*53*	*41*	*23*	*55*	*43*
9	*40*	*24*	*24*	*47*	*39*
10	*46*	*28*	*25*	*51*	*35*
11	*48*	*29*	*26*	*43*	*28*
12	*49*	*35*	*27*	*48*	*33*
13	*44*	*30*	*28*	*45*	*37*
14	*48*	*37*	*29*	*47*	*34*
15	*49*	*34*	*30*	*42*	*31*

Розв'язування. *Для здійснення даного групування скористаємося кростабуляцією, вибравши* ***Basic Statistics and Tables→Tables and banners.*** *Змінні виберемо як показано на рис. 3.9. Оскільки аналізуються тільки дві змінні, то буде побудована тільки одна таблиця. Таблиця кростабуляції зображена на рис.*

3.10. Щоб визначити відсоткове співвідношення, необхідно вибрати на вкладці ***Options*** *діалогового вікна* ***Crosstabulation Tables Results*** *позначку у полі відповідних відсотків, для того, щоб відсоткове співвідношення було подано в окремій таблиці (рис. 3.11). Також необхідно поставити позначку* ***Display Selected %'s in Sep. Tables*** *на вкладці* ***Advanced*** *діалогового вікна* ***Crosstabulation Tables Results*** *(рис. 3.12). Таблиця результатів зображена на рис. 3.13. Для графічного подання результатів необхідно вибрати або* ***категоризовані****, або* ***тривимірні гістограми*** *у вкладці* ***Advanced*** *діалогового вікна* ***Crosstabulation Tables Results*** *(рис. 3.12). Гістограма кростабуляції зображена на рис. 3.14. Якщо треба обчислити статистики (основні статистичні показники) слід вибрати необхідні показники, які будуть показані в* ***таблицях з двома змінними.***

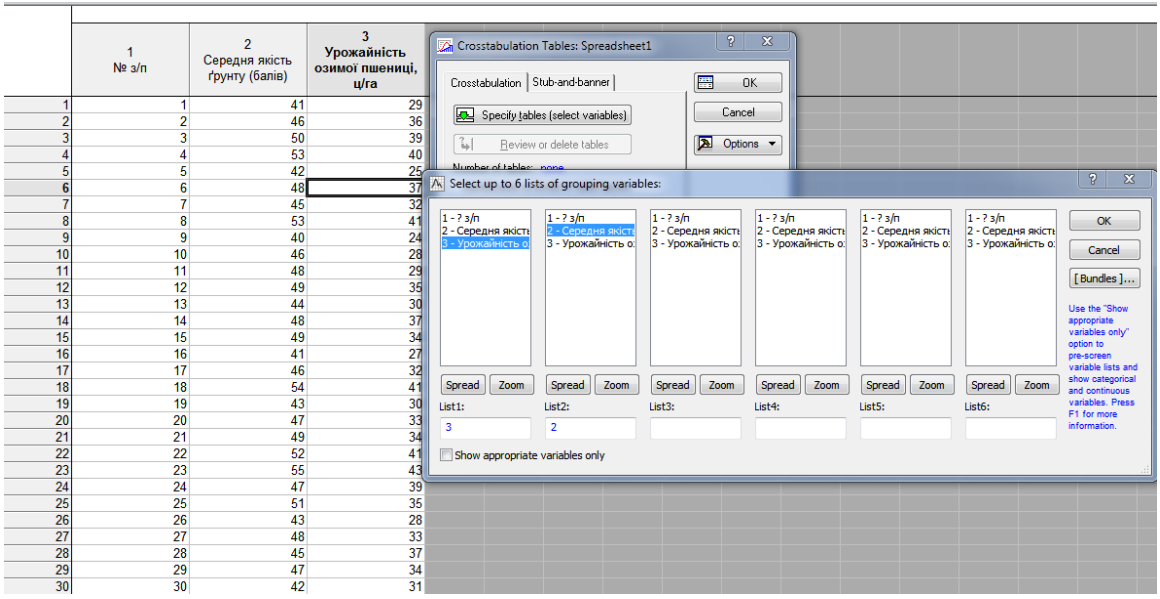


Рис. 3.9. Вибір змінних для групування за допомогою кростабуляції

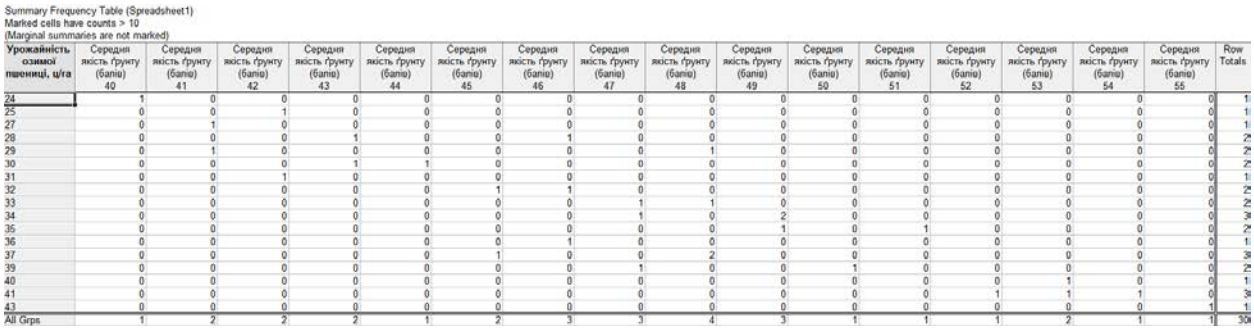


Рис. 3.10. Результати кростабуляції

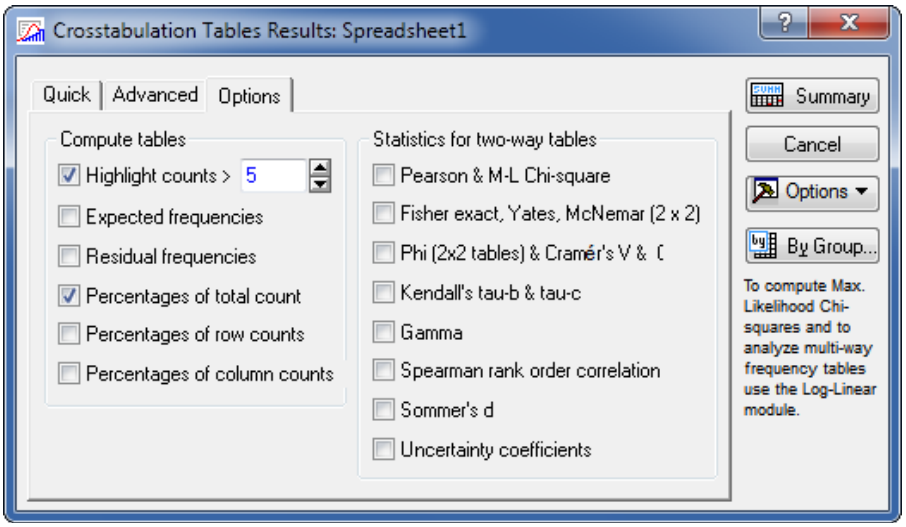


Рис. 3.11. Вибір параметрів у вкладці Options Crosstabulation Tables Results

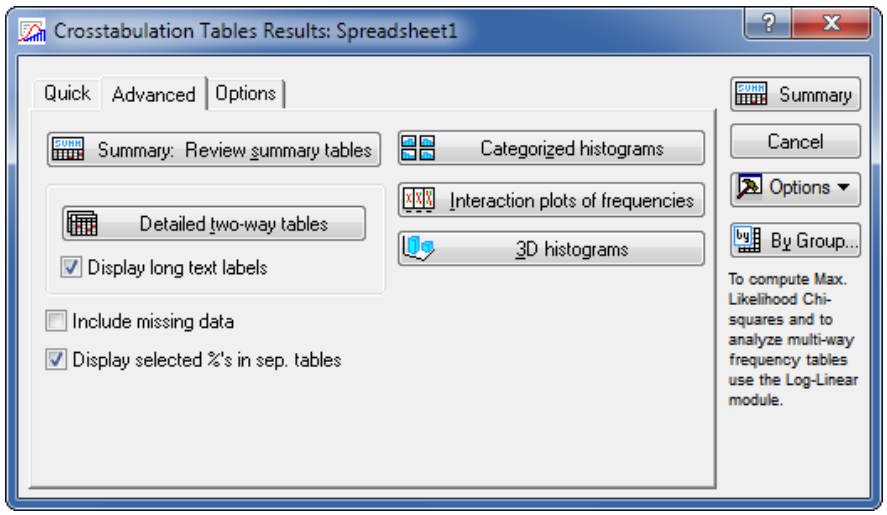


Рис. 3.12. Вибір способу подання відсоткового відображення результатів кростабуляції

Summary Table: Percentages of Total N=30 (приклад2)
Marked cells have counts > 10
(Marginal summaries are not marked)

Урожайність озимої пшениці, ц/га	Середня якість ґрунту (балів) 40	Середня якість ґрунту (балів) 41	Середня якість ґрунту (балів) 42	Середня якість ґрунту (балів) 43	Середня якість ґрунту (балів) 44	Середня якість ґрунту (балів) 45	Середня якість ґрунту (балів) 46	Середня якість ґрунту (балів) 47	Середня якість ґрунту (балів) 48	Середня якість ґрунту (балів) 49
24	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
25	0,00%	0,00%	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
27	0,00%	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
28	0,00%	0,00%	0,00%	3,33%	0,00%	0,00%	3,33%	0,00%	0,00%	0,00%
29	0,00%	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%
30	0,00%	0,00%	0,00%	3,33%	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%
31	0,00%	0,00%	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
32	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	3,33%	0,00%	0,00%	0,00%
33	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	3,33%	0,00%
34	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%	6,67%
35	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%
36	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%	0,00%	0,00%
37	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%	0,00%	6,67%	0,00%
39	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%	0,00%
40	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
41	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
43	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
All Grps	3,33%	6,67%	6,67%	6,67%	3,33%	6,67%	10,00%	10,00%	13,33%	10,00%

Рис. 3.13. Результат кростабуляції у відсотках

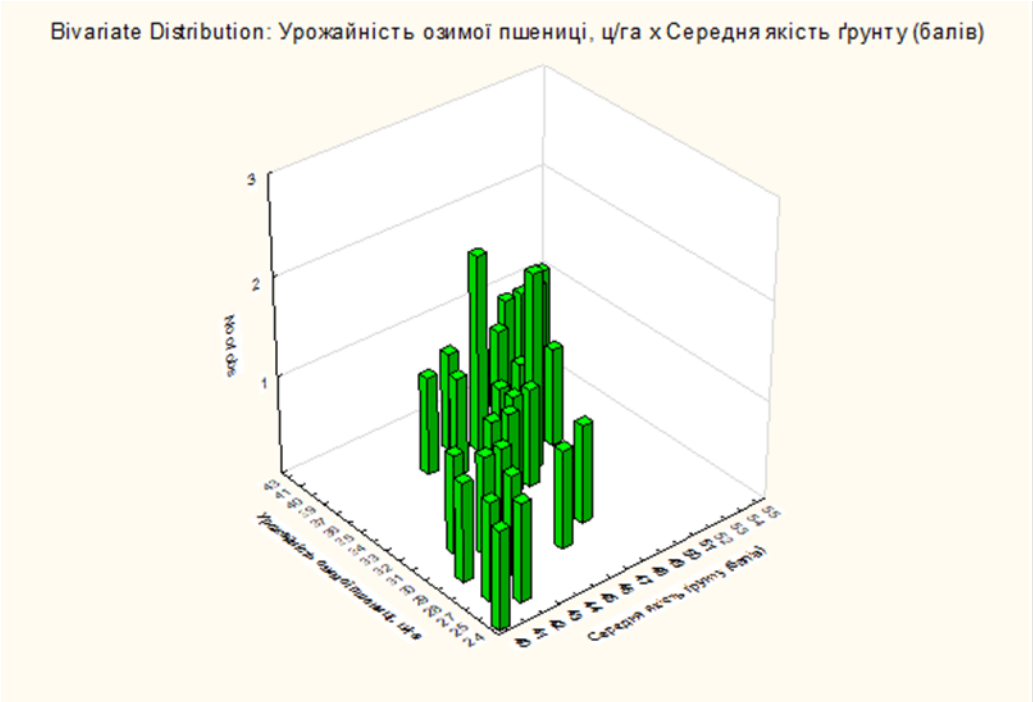


Рис. 3.14. Гістограма кростабуляції

Завдання для самостійної роботи

3.1. На основі вибіркової сукупності даних (табл. 3.2) про наявність захворювання серця (1 – наявність захворювання, 2 – відсутність захворювання) та факту паління (1 – палить; 2 – не палить) визначити на основі таблиць кростабуляції:

1) залежність наявності захворювання від факту паління;

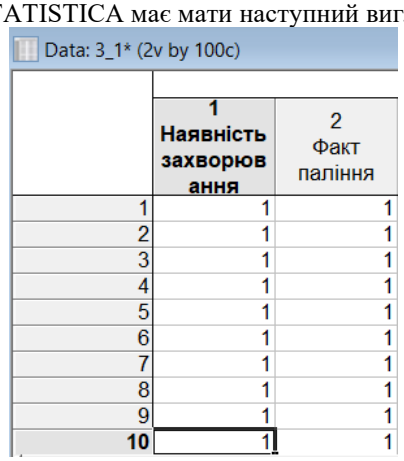
2) всі статистичні характеристики і пояснити їх зміст;

3) проілюструвати наявні залежності графічно.

Таблиця 3.2

№ з/п	*Наявність захворювання*	*Факт паління*	*№ з/п*	*Наявність захворювання*	*Факт паління*	*№ з/п*	*Наявність захворювання*	*Факт паління*
1	1	1	35	1	1	69	2	2
2	1	1	36	1	1	70	2	2
3	1	1	37	1	1	71	2	2
4	1	1	38	1	1	72	2	2
5	1	1	39	1	1	73	2	2
6	1	1	40	1	2	74	2	2
7	1	1	41	1	2	75	2	2
8	1	1	42	1	2	76	2	2
9	1	1	43	1	2	77	2	2
10	1	1	44	1	2	78	2	2
11	1	1	45	1	2	79	2	2
12	1	1	46	1	2	80	2	2
13	1	1	47	1	2	81	2	2
14	1	1	48	1	2	82	2	2
15	1	1	49	1	2	83	2	2
16	1	1	50	1	2	84	2	2
17	1	1	51	1	2	85	2	2
18	1	1	52	1	2	86	2	2
19	1	1	53	1	2	87	2	2
20	1	1	54	1	2	88	2	2
21	1	1	55	1	2	89	2	2
22	1	1	56	1	2	90	2	2
23	1	1	57	1	2	91	2	2
24	1	1	58	1	2	92	2	2
25	1	1	59	1	2	93	2	2
26	1	1	60	1	1	94	2	2
27	1	1	61	2	1	95	2	2
28	1	1	62	2	1	96	2	2
29	1	1	63	2	1	97	2	2
30	1	1	64	2	1	98	2	2
31	1	1	65	2	1	99	2	2
32	1	1	66	2	2	100	2	2
33	1	1	67	2	2			
34	1	1	68	2	2			

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх 100 спостережень.



3.2. На основі даних, наведених в таблиці 3.3, провести групувальний аналіз відносно переваг у використанні виду транспорту. Результати подати у частотах і відсотках.

Таблиця 3.3

№ з/п	*Категорія пасажирів*	*Вид транспорту*	*№ з/п*	*Категорія пасажирів*	*Вид транспорту*
1	Робітники	Автобус	21	Службовці	Метро
2	Службовці	Маршрутне таксі	22	Службовці	Автобус
3	Домогосподарки	Метро	23	Домогосподарки	Маршрутне таксі
4	Робітники	Автобус	24	Робітники	Маршрутне таксі
5	Робітники	Автобус	25	Домогосподарки	Автобус
6	Службовці	Метро	26	Службовці	Автобус
7	Домогосподарки	Маршрутне таксі	27	Домогосподарки	Метро
8	Службовці	Маршрутне таксі	28	Робітники	Маршрутне таксі
9	Службовці	Автобус	29	Робітники	Автобус
10	Домогосподарки	Метро	30	Робітники	Автобус
11	Домогосподарки	Маршрутне таксі	31	Робітники	Автобус
12	Службовці	Маршрутне таксі	32	Службовці	Маршрутне таксі
13	Робітники	Автобус	33	Робітники	Маршрутне таксі
14	Робітники	Маршрутне таксі	34	Домогосподарки	Метро
15	Домогосподарки	Автобус	35	Домогосподарки	Метро
16	Службовці	Автобус	36	Домогосподарки	Метро
17	Робітники	Автобус	37	Робітники	Метро
18	Домогосподарки	Метро	38	Робітники	Автобус
19	Робітники	Метро	39	Службовці	Маршрутне таксі
20	Робітники	Метро	40	Службовці	Маршрутне таксі

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх 40 спостережень.


Data: 3_2 (2v by 40c)

	1 Категорія пасажирів	2 Вид транспорту
1	Робітники	Автобус
2	Службовці	Маршрутне таксі
3	Домогосподарки	Метро
4	Робітники	Автобус
5	Робітники	Автобус
6	Службовці	Метро
7	Домогосподарки	Маршрутне таксі
8	Службовці	Маршрутне таксі
9	Службовці	Автобус
10	Домогосподарки	Метро

3.3. За результатами вибіркового обстеження умов життя домогосподарств у регіоні (табл. 3.4) здійснити групування домогосподарств за кількістю дітей та середнім місячним доходом. Результати групування подати у вигляді таблиці кростабуляції, сформувати відповідні висновки.

Таблиця 3.4

№ домогосподарства	*Кількість дітей до 15 років*	*Середній місячний дохід, грн.*	*Середні грошові витрати на душу, грн.*	*№ домогосподарства*	*Кількість дітей до 15 років*	*Середній місячний дохід, грн.*	*Середні грошові витрати на душу, грн.*
1	3	1675	258	16	2	2950	639
2	2	2446	540	17	0	1512	507
3	2	2172	456	18	2	2467	466
4	1	2517	561	19	1	2326	572
5	3	1390	249	20	1	2004	538
6	2	1464	335	21	3	2403	454
7	0	1526	496	22	2	2037	433
8	3	1485	262	23	1	1704	505
9	2	1950	377	24	1	1488	436
10	1	1496	374	25	2	1454	350
11	2	2475	441	26	2	1956	389
12	1	1076	292	27	1	1616	472
13	3	1735	307	28	3	2446	437
14	4	1625	230	29	1	2715	646
15	0	1654	641	30	2	3340	750

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх 30 спостережень.

Data: 3_3* (3v by 30c)

	1 Кількість дітей до 15 років	2 Середній місячний дохід, грн.	3 Середні грошові витрати на душу, грн.
1	3	1675	258
2	2	2446	540
3	2	2172	456
4	1	2517	561
5	3	1390	249
6	2	1464	335
7	0	1526	496
8	3	1485	262
9	2	1950	377
10	1	1496	374

3.4. За даними завдання 3.3 здійснити групування домогосподарств за кількістю дітей та грошовими витратами на душу. Результати групування подати у вигляді таблиці кростабуляції з відсотками, сформувати відповідні висновки.

3.5. За результатами завдання 3.3 здійснити групування домогосподарств за кількістю дітей (тільки 1 або 2 дитини; 2 або 3 дитини; 4 або 0 дітей) та середнім місячним доходом. Результати групування подати у вигляді таблиці кростабуляції та побудувати гістограми. Сформувати висновки.

Лабораторна робота № 4
Графічний метод подання статистичних даних в системі STATISTICA

1. Основні теоретичні відомості побудови двовимірних графіків

Таблична форма подання статистичного матеріалу не завжди дає змогу достатньо наочно і чітко відобразити загальну картину стану чи розвитку досліджуваного явища, дослідження структури та порівняння явищ, розкрити закономірності зв'язку статистичних показників або їхнього розподілу. У зв'язку з цим поряд із статистичними таблицями для розв'язання вищеописаних завдань застосовують графічний спосіб зображення статистичної інформації. При правильній побудові графіки характеризуються виразністю, доступністю, сприяють аналізу явищ, їх узагальненню і вивченню.

Побудова статистичних графіків є трудомістким процесом, тому для полегшення й прискорення побудови статистичних графіків важливо засвоїти техніку їх побудови з використанням персональних комп'ютерів. Сучасні ПК дають можливість не тільки оперативно, якісно і з мінімальними витратами праці і часу забезпечити високий рівень автоматичної побудови різних видів графічних зображень, але і виконати (це має особливо велике значення) різноманітні варіанти їх побудови. Великі можливості для автоматичної побудови різних видів графічних зображень статистичних даних має **STATISTICA**. Методика побудови окремих видів графіків, які використовуються для зображення статистичних даних, а також техніка їх побудови за допомогою **STATISTICA** докладно розглядатимуться далі.

У системі **STATISTICA** існують різні способи доступу до побудови графіків:

- за допомогою вкладки ***Graphs (Графіки)***;
- за допомогою контекстного меню, клацнувши правою кнопкою миші на комірці даних.

Вкладка ***Graphs*** має три групи: ***Common (Поширені)***, ***More… (Додаткові)*** та ***Tools (Інструменти)*** (рис. 4.1).

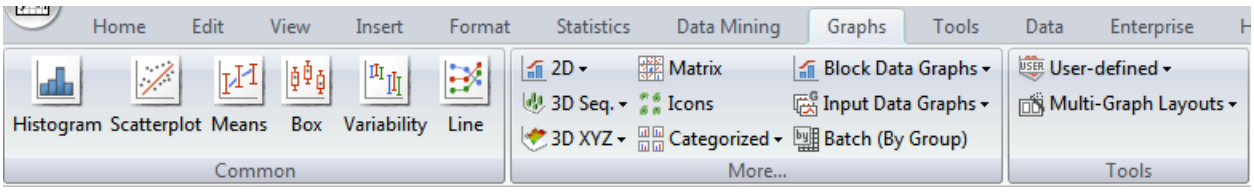


Рис. 4.1. Вкладка ***Graphs***

озглянемо деякі види статистичних двовимірних графіків (2D) (група ***Common*** або група ***More…***меню ***2D***):

- ***2D Histogramms (Двовимірні гістограми)*** – побудова гістограми;
  - ✓ ***2D Scatterplots (Діаграми розсіювання)*** – візуалізація залежності між двома вибраними змінними;
  - ✓ ***2D Box Plots (Діаграми розмаху)*** – побудова діаграми розмаху. Діапазони або характеристики розподілу обраної змінної (змінних) зображуються окремо для кожної групи спостережень. Для кожної групи спостережень обчислюється центральний момент (наприклад, медіана або середнє) і варіаційні статистики або статистики діапазону (наприклад, квартилі, стандартні помилки або середньоквадратичне відхилення). Програма навколо середньої точки зображає прямокутник і відрізки, що відображають діапазон розкиду;
  - ✓ ***2D Line Plots (Лінійні графіки)*** – побудова лінійних графіків;

- ✓ ***2D Bar/Column Plots (Стовпчикові діаграми)*** – подання послідовності значень у вигляді стовпців (одному спостереженню відповідає один стовпчик);
- ✓ ***Pie Charts (Кругова діаграма)*** – побудова кругової діаграми.

Технологію побудови вказаних вище статистичних графіків розглянемо на основі даних типового прикладу лабораторної роботи № 2. Для побудови гістограми необхідно з групи ***More… (Додаткові)*** вибрати меню ***2D (2D графіки)→ (Гістограми)*** (рис. 4.2) або кнопку ***Histogram*** групи ***Common***. З'явиться діалогове вікно ***2D Histogramms*** (рис. 4.3).

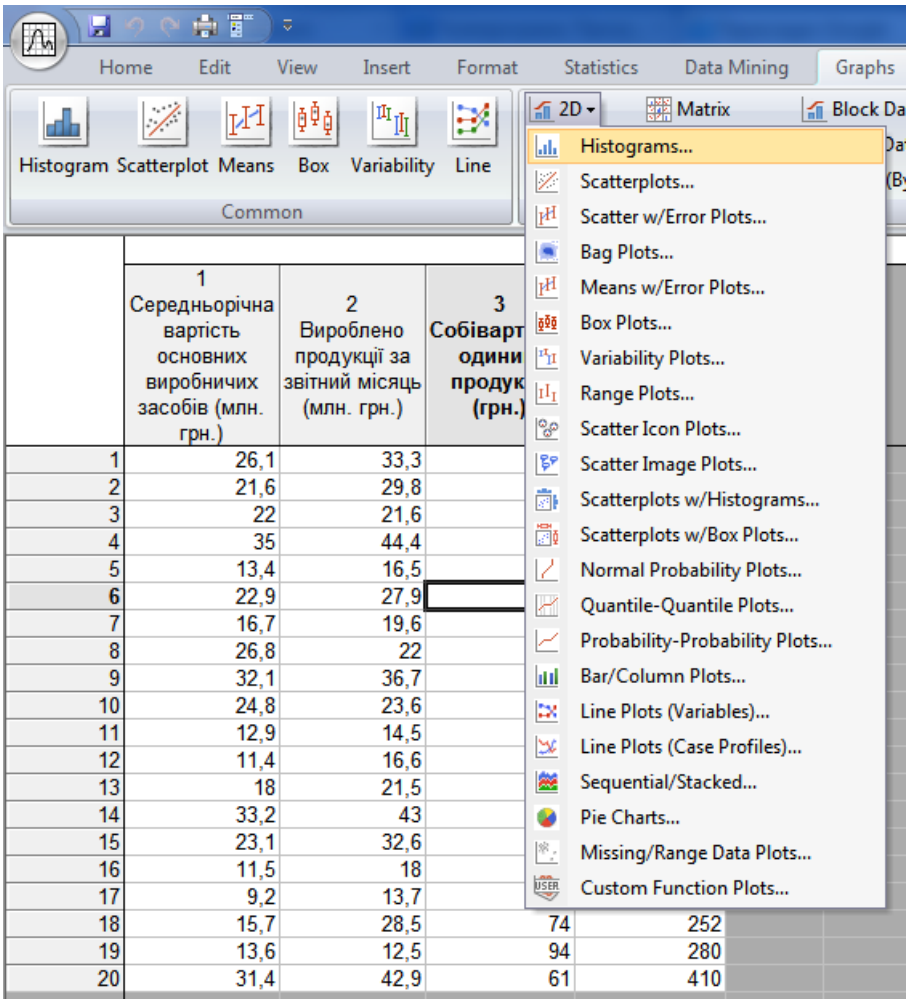


Рис. 4.2. Вибір графіку

На вкладці ***Advanced*** діалогового вікна ***2D Histograms*** (рис. 4.3) в полі ***Graph type (Тип графіка)*** вказується тип графіка: ***Regular (Простий)***, ***Multiple (Складений)*** та ***Double-Y (З подвійною віссю Y)***. У полі ***Fit type (Тип підгонки)*** вибираються види апроксимуючих законів розподілів: ***Off (Вимкнути)***, ***Normal (Нормальний)***, ***Beta (Бета)***, ***Exponential (Експоненційний)*** тощо. У списку ***Showing type (Тип показу)*** вказуються формати графіків: ***Standard (Стандартний)***, ***Hanging Bars (Висячі прямокутники)***, ***Cumulative (Кумулятивний)*** та ***Stacked (Складений)***.

У рамці ***Intervals (Інтервали)*** діалогового вікна рис. 4.3 встановлюються режими категоризації. У режимі ***Integer mode (Цілі числа)***, якщо не встановлений прапорець в полі ***Auto***, програма округлить кожне значення виділеної змінної до цілого числа і створить одну категорію (або графік у випадку категоризованих графіків) для кожного цілочислового значення. При виборі цього методу кнопка ***Change variable (Змінити змінну)*** дозволить вибрати іншу змінну. Якщо число цілих категорій перевершить 256, програма автоматично

використовує метод категоризації, що містить 16 категорій. У полі введення праворуч від режиму ***Categories (Категорії)*** вводиться необхідне число категорій. Програма розділить повний діапазон значень змінної на задане число інтервалів однакової довжини (довжина інтервалів не буде цілим числом). Після вибору опції ***Boundaries (Межі)*** треба натиснути кнопку ***Specify Boundaries (Задати межі)*** та ввести список меж для виділеної змінної в діалоговому вікні. Процедура можлива, якщо в полі ***Fit type*** обрано режим ***Off***. Опцію ***Codes (Коди***) можна використовувати, якщо змінна містить коди, за якими потрібно задати категорії. Після вибору цієї опції треба натиснути кнопку ***Specify Codes (Задати коди)*** і ввести потрібні коди в діалоговому вікні. Це можливо, якщо в полі ***Fit type*** обрано режим ***Off***. Після вибору методу ***Multiple subsets (Складені підгрупи)*** треба натиснути кнопку ***Specify subsets (Задати підгрупи)*** і у вікні задати умови вибору. Цей метод дозволяє використовувати більше однієї змінної для визначення груп. У рамці ***Statistics (Статистики)*** вибираються критерії відповідності емпіричних розподілів розподілам, наведеним у полі ***Fit type***. Якщо всі параметри гістограми вибрані, то відповідно натискають ***OK***.

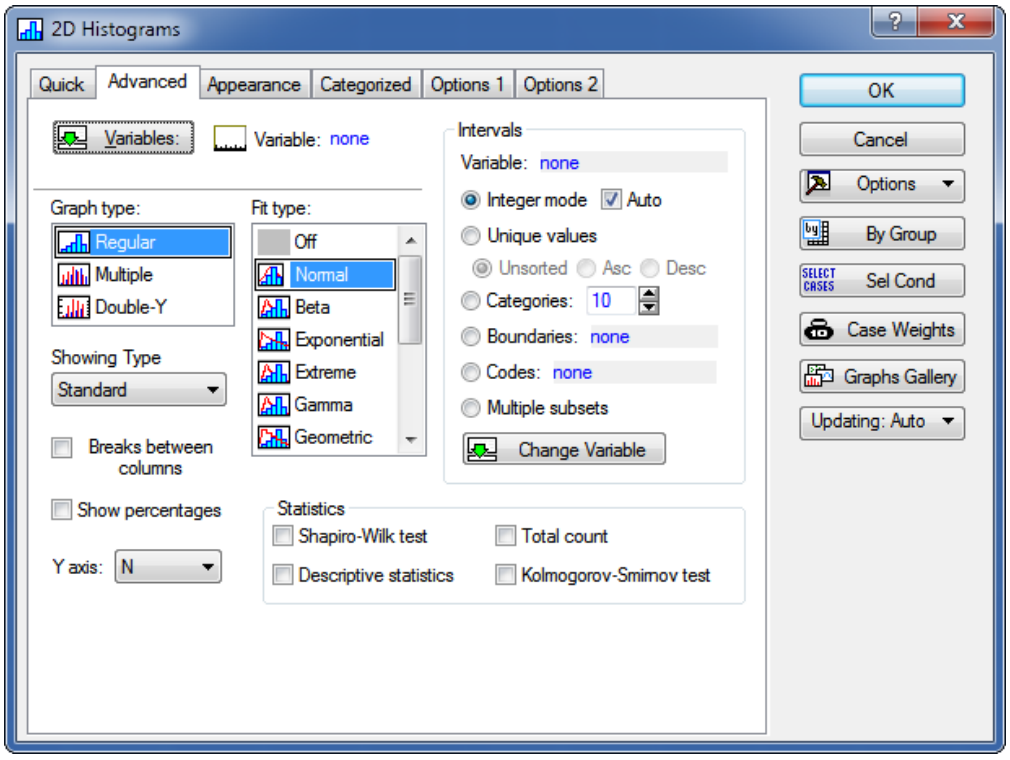


Рис. 4.3. Вкладка ***Advanced*** діалогового вікна ***2D Histograms***

Для побудови діаграми розсіювання необхідно з групи ***More… (Додаткові)*** вибрати меню ***2D→ Scatterplots (Діаграми розсіювання)*** або кнопку ***Scatterplot*** групи ***Common*** (рис. 4.1). З'явиться діалогове вікно ***2D Scatterplots*** (рис. 4.4).

На вкладці ***Advanced*** в полі ***Graph type*** можна вибрати типи діаграм: ***Regular***, ***Multiple***, ***Double-Y***, ***Frequency (розсіювання частот)***, ***Bubble*** (***розсіювання бульбашок***), ***Quartile (розсіювання квартилів)*** та ***Voronoi (Вороного)***.

У полі ***Fit*** діалогового вікна рис. 4.4 можна вибрати функції, які будуть описувати залежність вибраних двох змінних. Можливий вибір наступних функцій: ***Linear (Лінійна)***, ***Polynomial (Поліноміальна)***, ***Logarithmic (Логірифмічна***), ***Exponential (Експоненційна)***, ***Distance Weighted LS (Відстань зважених найменших квадратів)***, ***Neg Expon Weighted LS (Від'ємна експоненційна зважена)***, ***Spline (Бікубічне згладжування)*** та ***Lowess (згладжування пари даних)***.

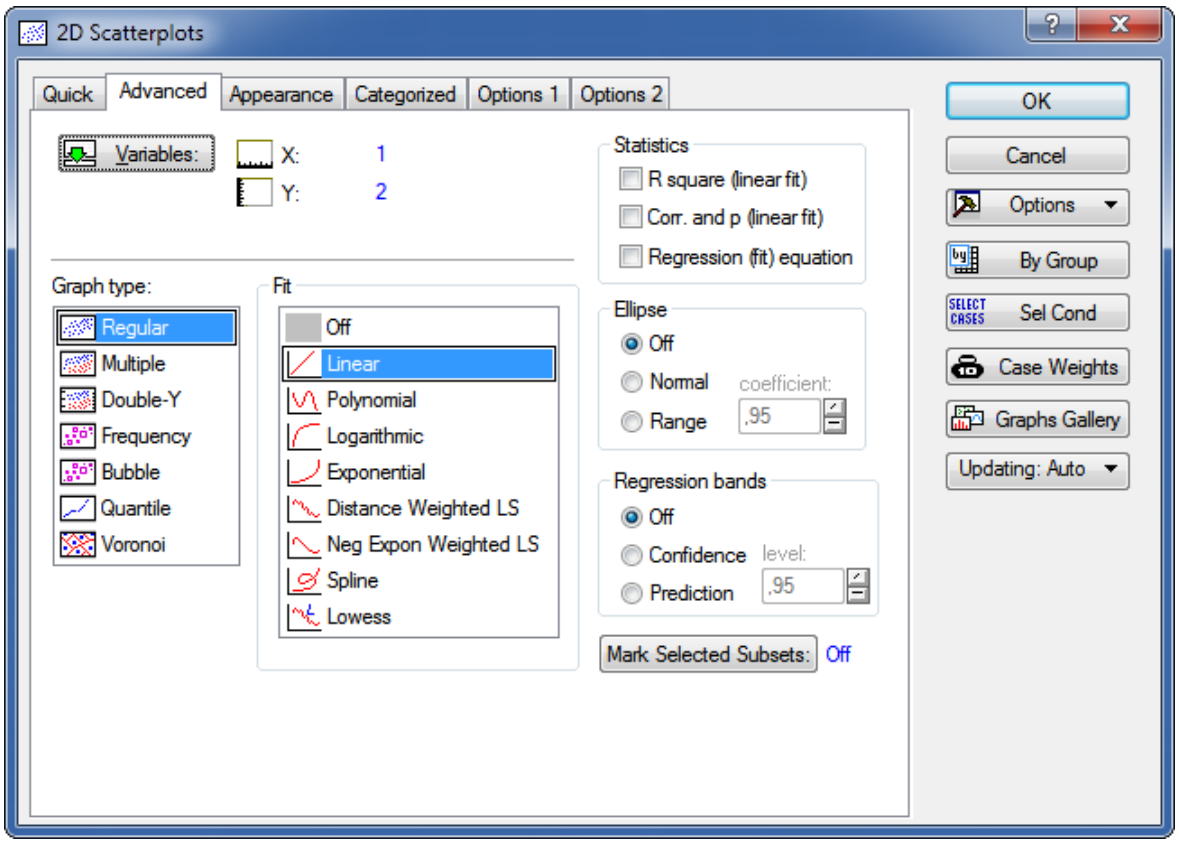


Рис. 4.4. Діалогове вікно ***2D Scatterplots***

У рамці ***Ellipse*** (***Елипс***) діалогового вікна рис. 4.4 опція ***Normal*** задає побудову еліпса в припущенні про нормальний розподіл двовимірної випадкової величини (X, Y), де, наприклад, X – змінна v1, Y – змінна v2. Орієнтація еліпса визначається знаком лінійної кореляції між двома змінними. Еліпс показує довірчий інтервал для одного спостереження за даних оцінок параметрів двовимірного нормального розподілу. Якщо кількість спостережень незначна, то еліпс може вийти за межі області, показаної на графіку. Опція ***Range (Розмах)*** означає побудову еліпса фіксованого розміру. Опція ***Regression bands (Межі регресії)*** застосовується для лінійної або поліноміальної функції. У рамці ***Statistics*** вибираються статистичні характеристики залежності між змінними: ***R square (Коефіцієнт детермінації)***, ***Corr. and p (Коефіцієнт кореляції та рівень значущості p)*** та ***Regression (fit) equation (Рівняння регресії)***.

Статистичний графік ***2D Box Plots (Діаграма розмаху)*** має деякі особливості. Крім зазначених вище налаштувань, необхідно вибрати статистики для оцінки розкиду залежної змінної (рис. 4.5): ***Std.dev. (Середньоквадратичне відхилення)***, ***Std.error (Стандартна помилка)***, ***Conf.Interval (Довірчий інтервал)***, ***Min-Max (Мінімальне-максимальне)***, ***Constant (Константа)*** для прямокутника і відрізків та ***Non Outlier range (без викидів)*** додатково для відрізків. Ці статистики відповідають значенням ***Mean (Середнє)*** для центральної точки. При зміні статистики оцінки середнього на ***Median (Медіана)*** міняються статистики оцінки розкиду в зазначених полях. У полях з'являться оцінки ***Percentiles (Відсотків)***. Там же, в цих полях, вказуються коефіцієнти перед цими статистиками. У рамці ***Outliers (Викиди)*** задаються режими обробки викидів: ***Off (Вимкнути)***, ***Outliers (Викиди)***, ***Extremes (Крайні точки)***, ***Outliers & Extremes (Викиди і крайні точки)***.

Аналогічно задаються параметри інших статистичних графіків в системі **STATISTICA**.

2. Додаткові налаштування побудови статистичних графіків

Розглянемо приклад додаткового налаштування описаних вище графіків. Спершу побудуємо декілька лінійних графіків на основі даних типового прикладу лабораторної роботи № 2. Відкривши файл з даними, з групи ***More… (Додаткові)*** виберемо меню ***2D (2D графіки)→Line Plots (Variables) (Лінійні графіки (для змінних))*** (рис. 4.1). З'явиться діалогове вікно ***2D Line Plots*** (рис. 4.6).

Виберемо три змінні для побудови статистичного лінійного графіку. У полі ***Graph type (Тип графіка)*** вкладки ***Advanced*** діалогового вікна ***2D Line Plots*** наведений список доступних для побудови лінійних графіків. Виділяють декілька типів лінійних графіків (табл. 4.1).

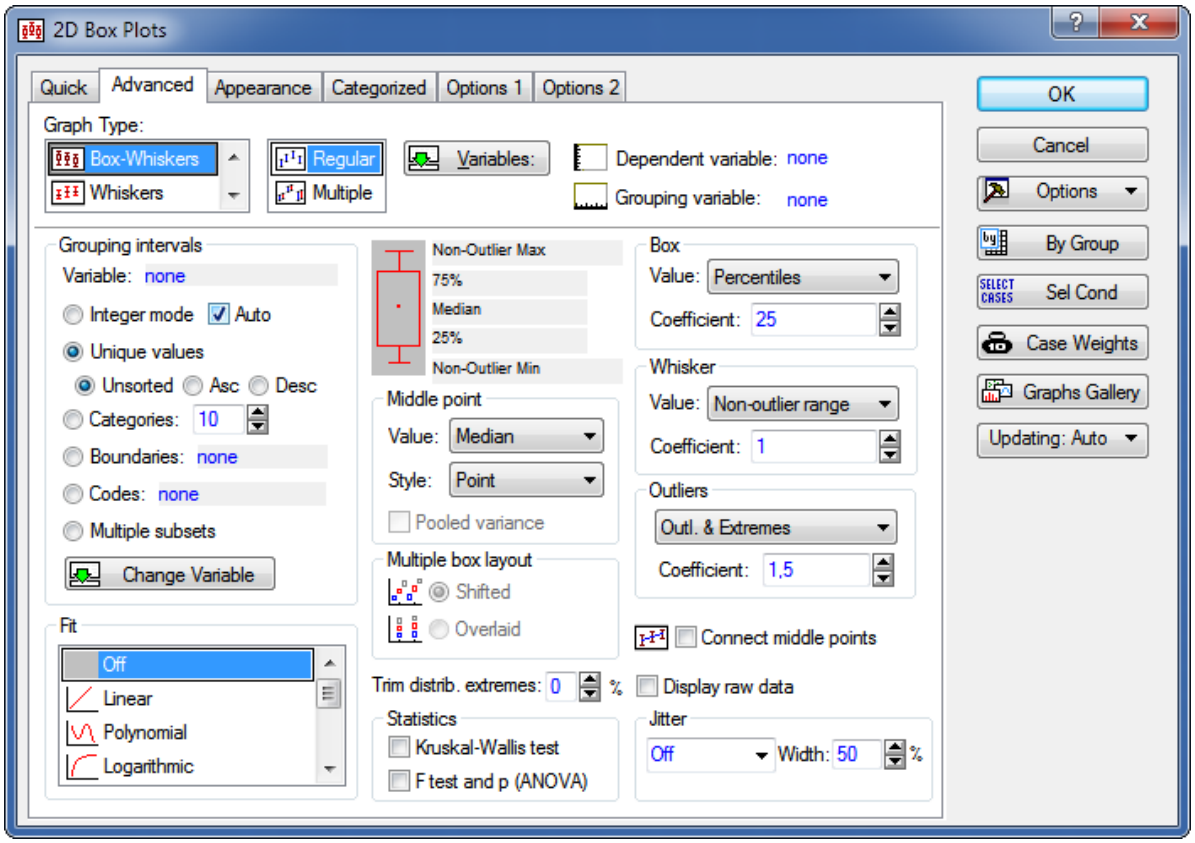


Рис. 4.5. Діалогове вікно ***2D Box Plots*** (вкладка ***Advanced***)

За замовчуванням вибирається простий лінійний графік однієї змінної. Якщо просто натиснути ***OK***, то для кожної із змінних буде побудовано окремий графік, в окремому вікні. Оскільки потрібно зобразити всі три змінні на одному графіку, то в діалоговому вікні ***2D Line Plots*** необхідно вибрати тип ***Multiple.*** Графік матиме наступний вигляд (рис. 4.7).

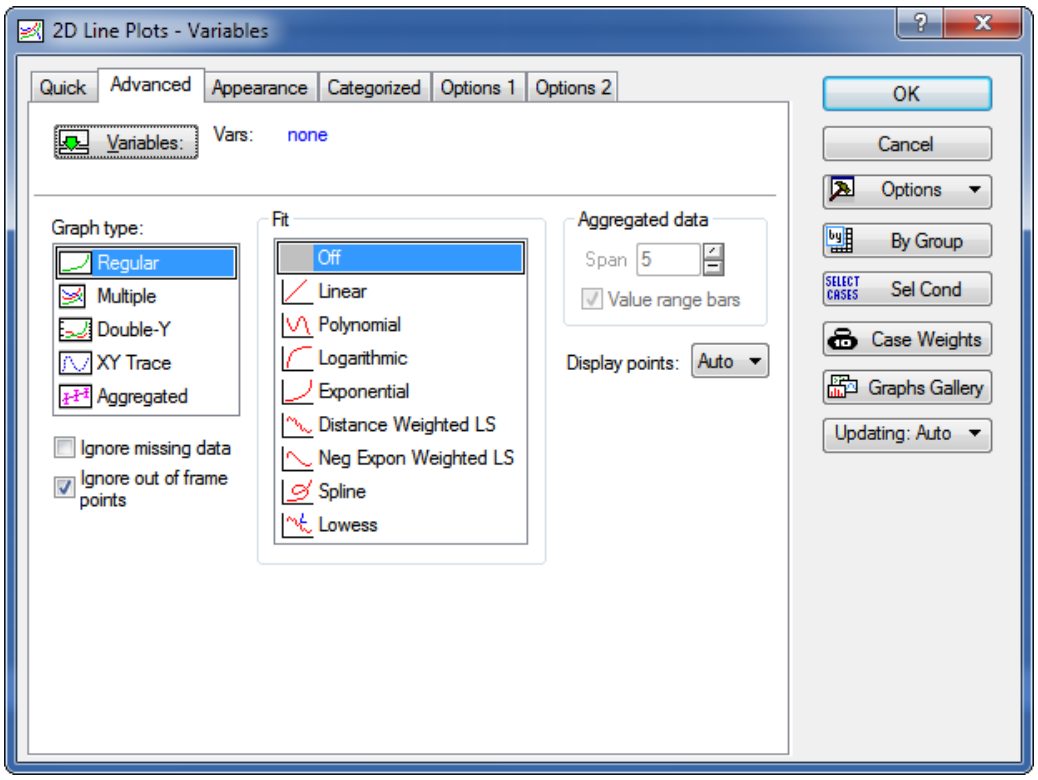


Рис. 4.6. Діалогове вікно ***2D Line Plots*** (вкладка ***Advanced***)

Таблиця 4.1

Типи графіків (***Graph type***)

Тип графіку	Особливості
Regular	Крива, яка з'єднує ряд значень змінної. Порожня комірка даних (тобто пропущені дані) «розриває» лінію.
Double-Y	Комбінація двох по-різному масштабованих лінійних графіків. Для кожної обраної змінної використовується свій шаблон лінії. Лінійний графік з подвійною віссю Y (ординат) можна використовувати для порівняння рядів значень кількох змінних, накладаючи їх лінійні подання на один графік. У той же час в силу незалежності шкал, що використовуються для двох осей, цей графік може полегшити зіставлення змінних, що важко піддаються порівнянню (тобто мають значення в різних діапазонах).
Multiple	Використовується для зображення кількох графіків в одному вікні. Для кожної змінної використовується і вказується в умовних позначеннях свій шаблон і колір лінії. Цей тип лінійних графіків використовується для порівняння значень кількох змінних (або кількох функцій) шляхом зображення їх на одному графіку.
XY Trace (Трасувальний)	Спочатку будується діаграма розсіювання двох змінних, а потім окремі точки даних з'єднуються лінією (в порядку їх зчитування з файлу даних).
Aggregated (агрегований)	Графіки, які зображують ряди середніх для послідовних підмножин обраної змінної. Можна вибрати число послідовних спостережень, за якими буде обчислено середнє, а при необхідності діапазон значень в кожному підмножині буде виділений значками типу відрізків. Агреговані лінійні графіки використовуються для подання та дослідження послідовностей великого масиву значень.

Редагувати графіки можливо на вкладці ***Edit (Правка)*** головного меню у відповідних блоках. Однак, існують і способи швидкої зміни елементів графіка, які не вимагають виконання великої кількості дій. Два основні правила редагування графіків наступні:

1) для вибору конкретного способу редагування об'єкту потрібно клацнути правою кнопкою миші на цьому об'єкті і вибрати тип редагування з контекстного меню;

2) щоб отримати доступ до загальних (встановлених за замовчуванням) способів редагування об'єкту, двічі клацнути на вибраному об'єктів. Наприклад, щоб змінити назву графіку потрібно двічі клацнути мишею в його полі, ввести нове ім'я і натиснути ***ОК***.

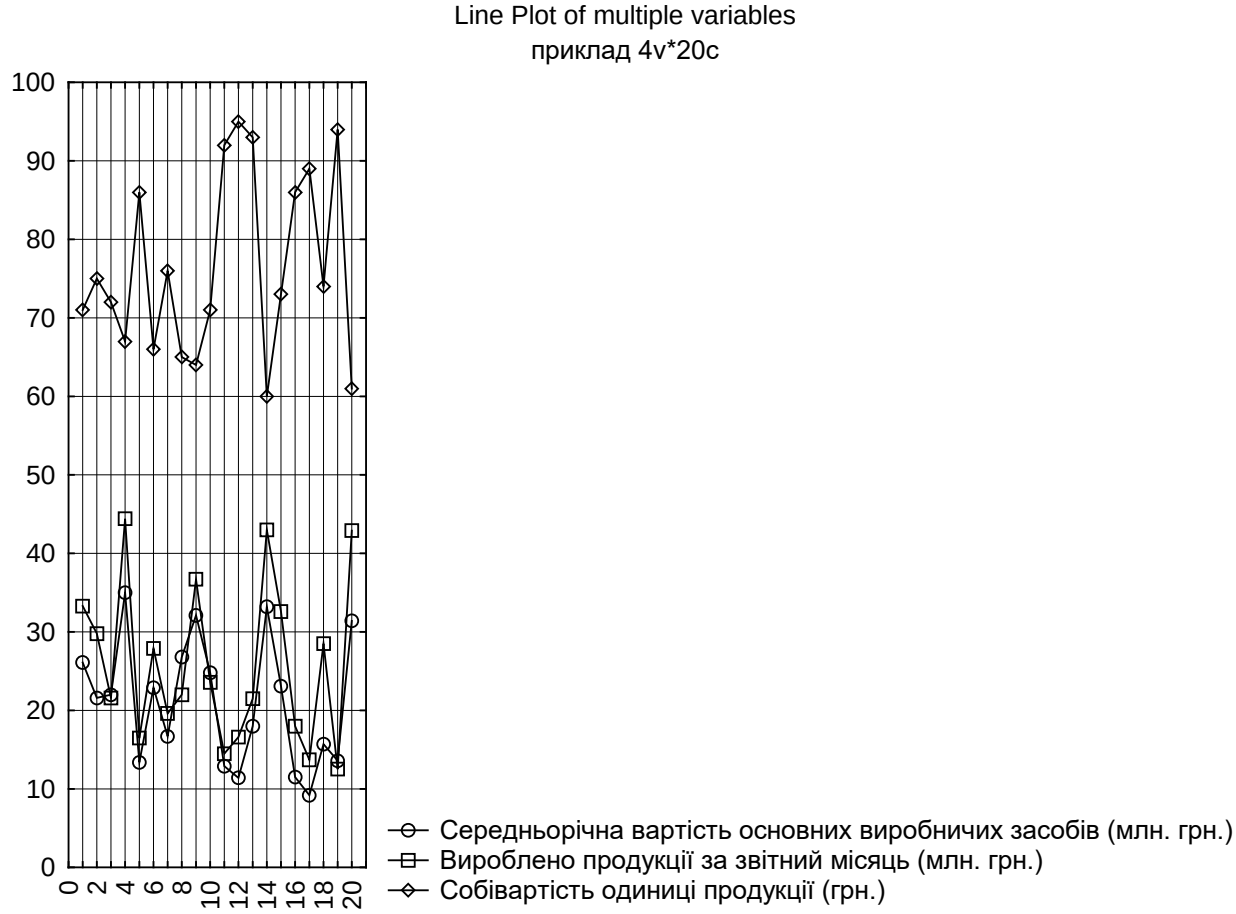


Рис. 4.7. Лінійний графік для багатьох змінних

Можливий також інший спосіб редагування: зробивши подвійне клацання на фоновій поверхні вікна, можна викликати діалогове вікно ***Graph Option (Опції графіку)***, в якому зведені всі режими редагування, в тому числі і заголовків (рис. 4.8).

Як видно з рис. 4.7, легенда може бути розташована невдало. Щоб змінити розташування легенди чи будь-якого іншого надпису необхідно в діалоговому вікні ***Graph Option*** вибрати ***Titles/Text (Заголовки/Текст)***, далі вибрати необхідне поле для редагування (рис. 4.9 а) та вказати статус або тип розміщення у списку ***Status (Статус)*** (рис. 4.9 б), де можна обрати ***Title (Заголовок)***, ***Subtitle (Підзаголовок)***, ***Footnote (Примітка)***, ***Left (Зліва)***, ***Right (Справа***) та ***Floating (Рухомий***). Якщо вказати тип розміщення ***Floating***, текст буде поданий в рамці, яку можна перемістити в будь-яке місце графіку.

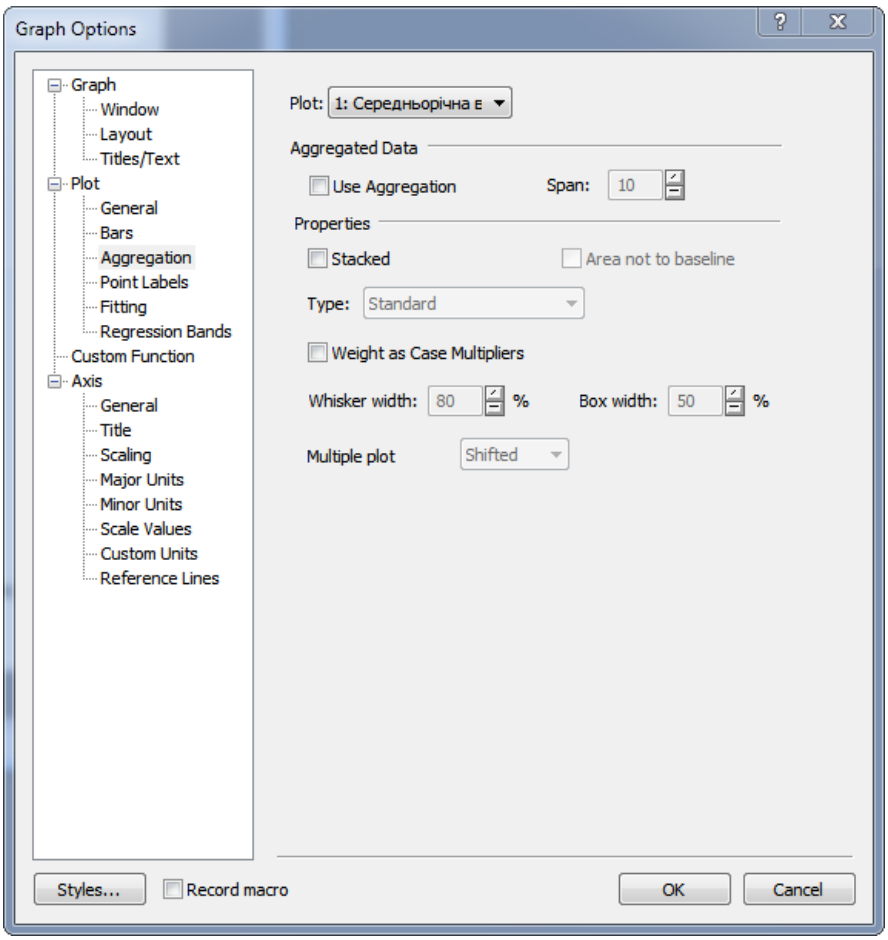


Рис. 4.8. Діалогове вікно редагування графіку

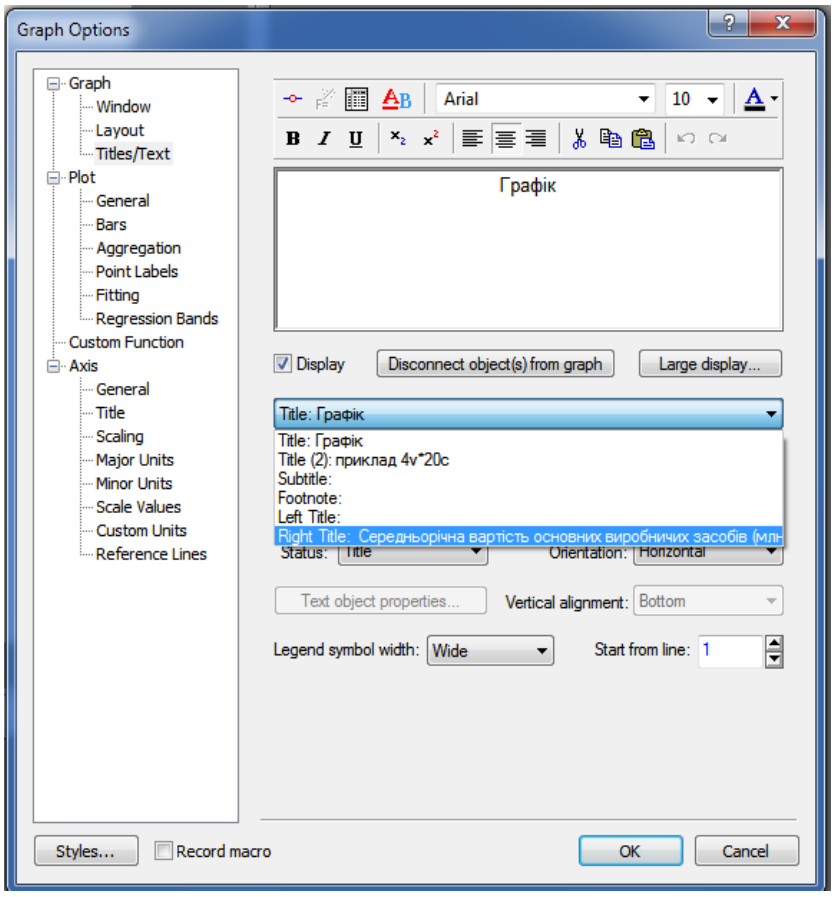


Рис. 4.9 а). Вибір поля для редагування параметрів графіку

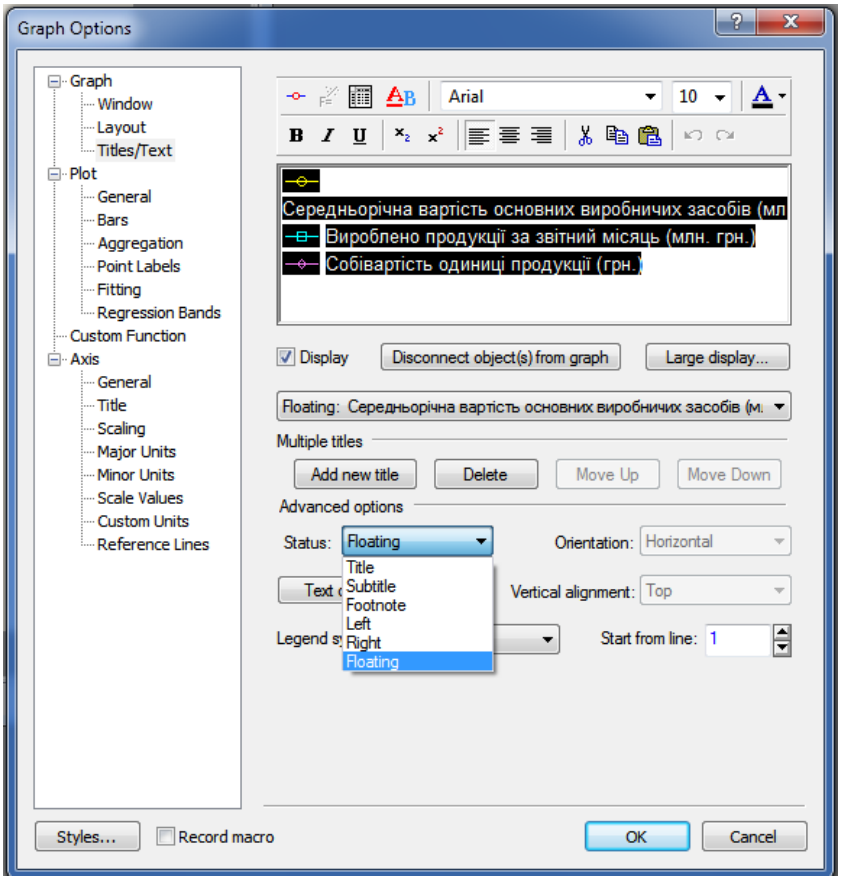


Рис. 4.9 б). Редагування розміщення легенди графіка

Для кожного графіку масштаб може бути підібраний окремо і вказаний на лівій або правій осі ординат. Можна добитися “кращої подачі” змінної, якщо встановити для неї окремий масштаб уздовж правої осі ординат, включивши при цьому автоматичний режим оптимального масштабування. Наприклад, змінна собівартість одиниці продукції розміщена значно вище, ніж інші графіки. Можна розташувати її вздовж правої осі ординат. Для цього двічі клацнемо на графіку змінної та виберемо в діалоговому вікні ***Graph Option*** у полі ***Plot (Графік)→ General (Загальні)*** розташування вздовж правої осі Y (рис. 4.10 а).

Зміна призначення осей приводить до зміни графіку. Графік змінної собівартість одиниці продукції став більше розтягнутим уздовж осі Y. Оскільки ця залежність побудована тепер уздовж правої осі Y, то на цій осі повинні бути і відповідні позначення. Для зміни позначень необхідно зробити подвійне клацання на правій осі Y. У діалоговому вікні ***Graph Option*** вибрати ***Axis (Вісь)→General (Загальні)***, потім натиснути кнопку ***Summary of settings…(Огляд налаштувань)*** і встановити прапорець у групі ***Scale values (Шкала значень)*** біля одного із варіантів: ***Automatic (Автоматична)***, ***Variable values (Значення змінних)*** та ***Text labels (Текстові позначення)*** (рис. 4.10 б).

Масштабування осей відбувається також у діалоговому вікні ***Graph Option (Axis→Scaling (Масштабування))*** (рис. 4.11 а). Передбачено два режими розмітки осі: ***Auto (Автоматичний)*** і ***Manual (Ручний)***. Якщо вибрана розмітка ***Auto***, то програма сама вибирає мінімальне і максимальне значення на шкалі таким чином, щоб всі точки на графіку були показані. Якщо вибрати режим ***Manual***, то параметри ***Minimum (Мінімум)***, ***Maximum (Максимум)*** та ***Edit step (Редагувати крок)*** визначаються користувачем. Якщо необхідно задати формати подання позначень осі, слід перейти в меню ***Axis→Scale values (Шкала значень)*** та вибрати необхідний формат (рис. 4.11 б).

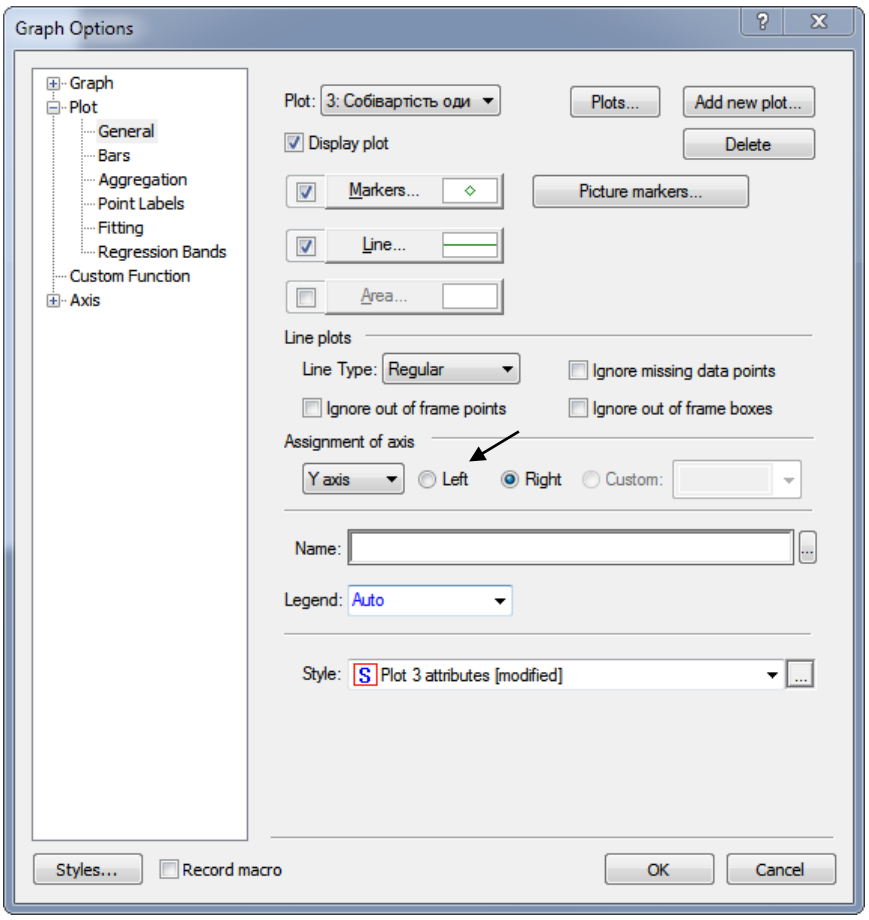


Рис. 4.10 а). Вибір осі побудови графіка

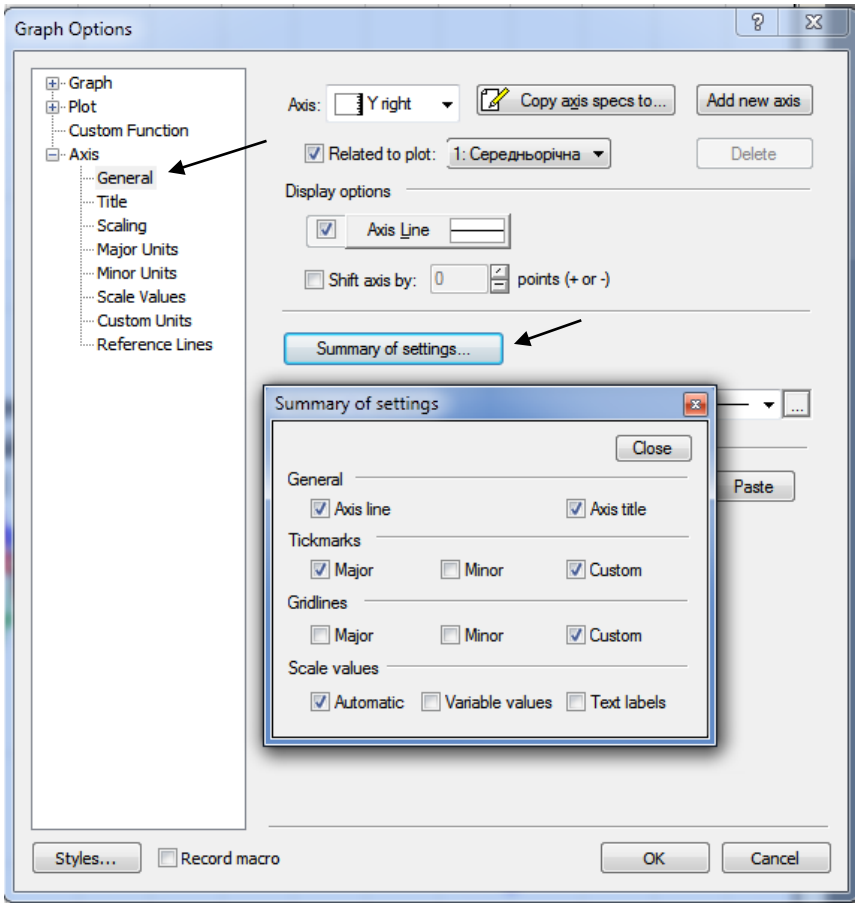


Рис. 4.10 б). Налаштування вісі графіка

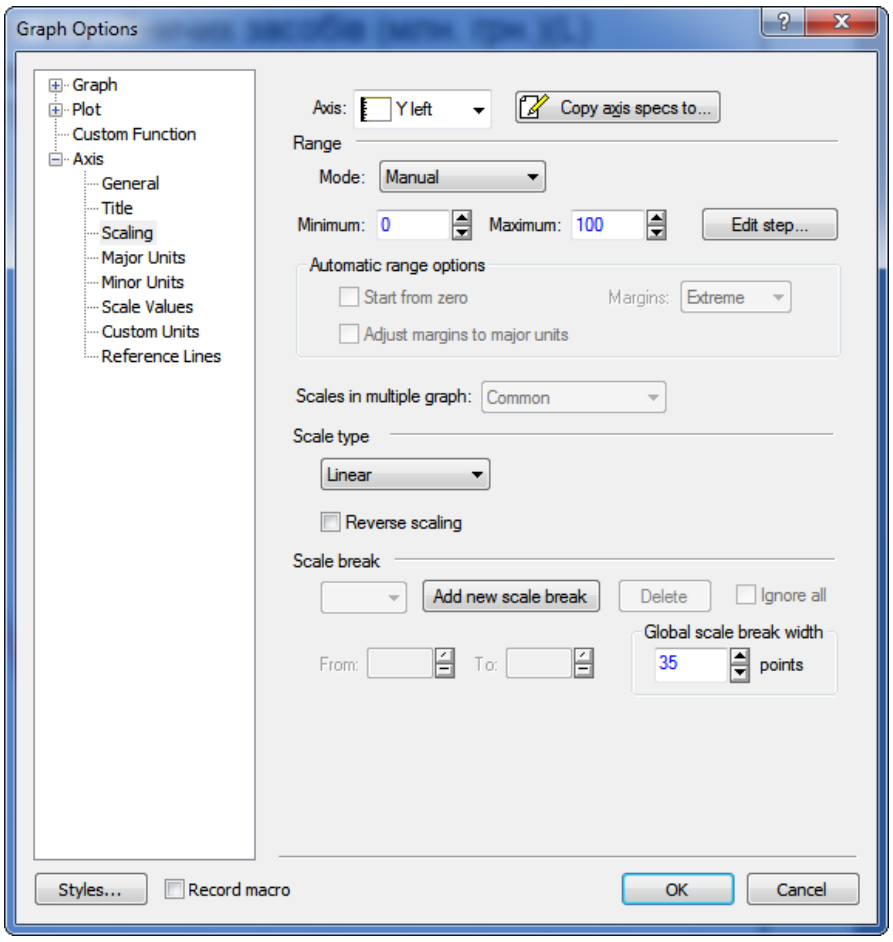


Рис. 4.11 а). Масштабування осей

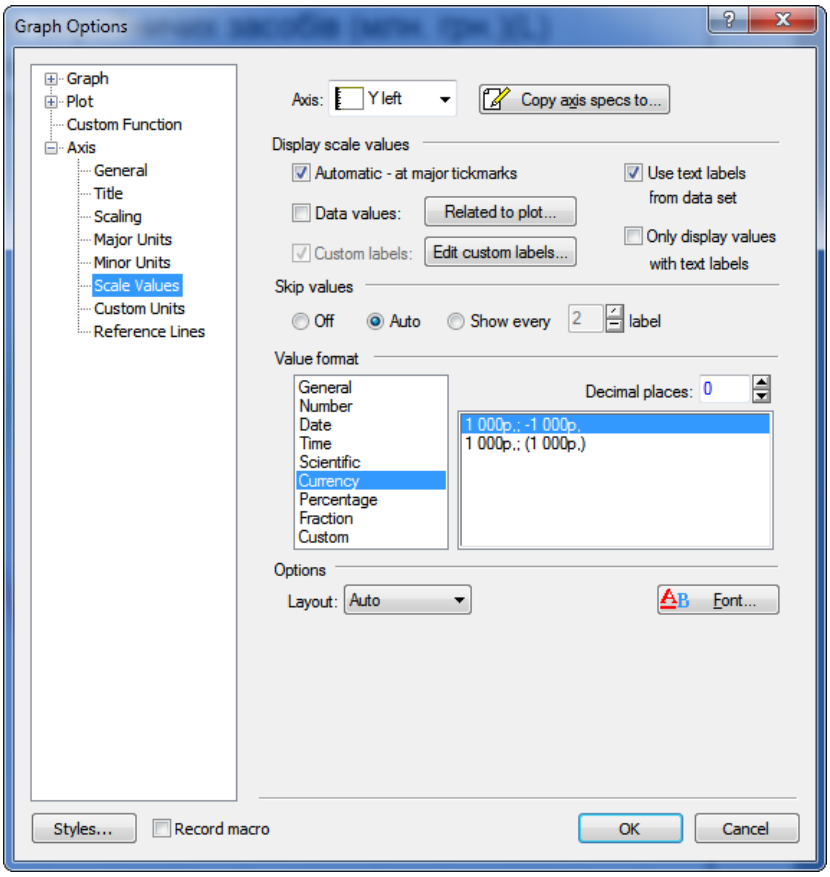


Рис. 4.11 б). Задання форматів подання позначень осі

Часто необхідно показати, що позиція яка приймається за 0 зовсім не відповідає нульовій відмітці на графіку. Це можна зробити, ввівши розрив шкали на даній осі. Щоб ввести розрив шкали для осі, слід перейти в меню ***Axis→Scaling*** і вибрати ***Add New scale break (Додати розриви у масштабі)***, після чого вказати значення розриву (рис. 4.12).

У графіках можна також змінити тип подання графіку. Для цього в діалоговому вікні ***Graph Option (Plot→General)*** вибрати необхідний тип подання у списку ***Style (Стиль)*** (рис. 4.13).

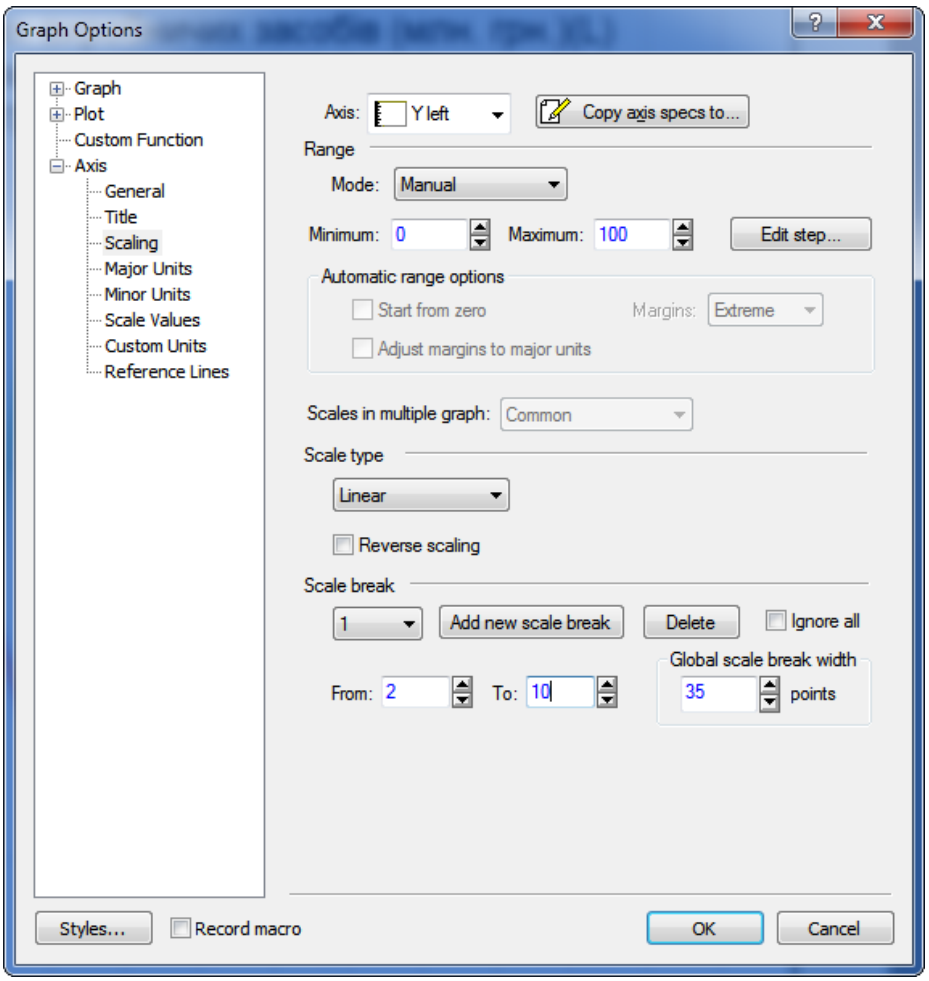


Рис. 4.12. Встановлення розривів на осі

При вивченні зв'язку між змінними за допомогою діаграм розсіювання (***2D Scatterplots***) у **STATISTICA** є можливість вивчати та виявляти коливання ("викиди"), які, як правило, притаманні аномальним або нетиповим даним. Суттєві коливання вливають на основні кореляційні характеристики досліджуваного зв'язку, наприклад коефіцієнт кореляції. **STATISTICA** надає можливість видалити суттєві коливання значень змінної за допомогою інструменту ***Brushing (Зафарбовування)***. При цьому, якщо видалити аномальні значення змінної, функція регресії буде мати інший вигляд.

Розглянемо основні принципи роботи із зафарбовуванням. Графік, побудований для прикладу, зі всіма змінами матиме вигляд, як на рис. 4.14. Щоб добавити на графік лінію тренду, необхідно в діалоговому вікні ***Graph Option (Plot (Графік)→ Fitting (Підгонка))*** натиснути кнопку ***Add new fit (Додати нову лінію тренду)***.

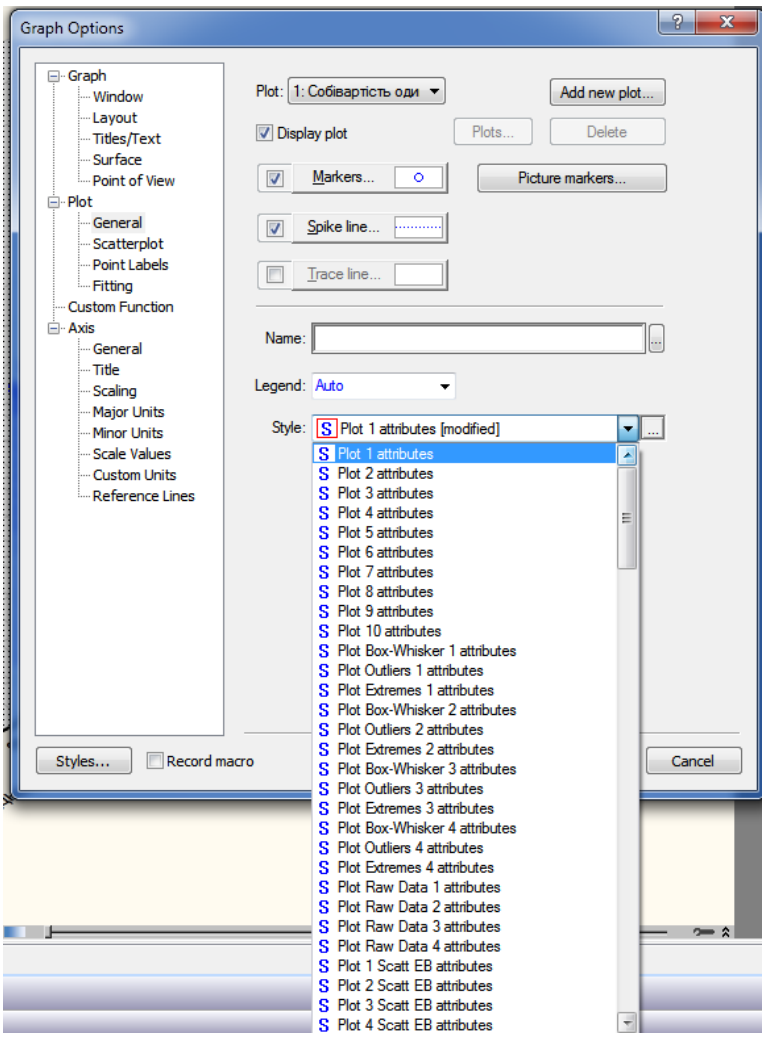


Рис. 4.13. Зміна типу подання залежності

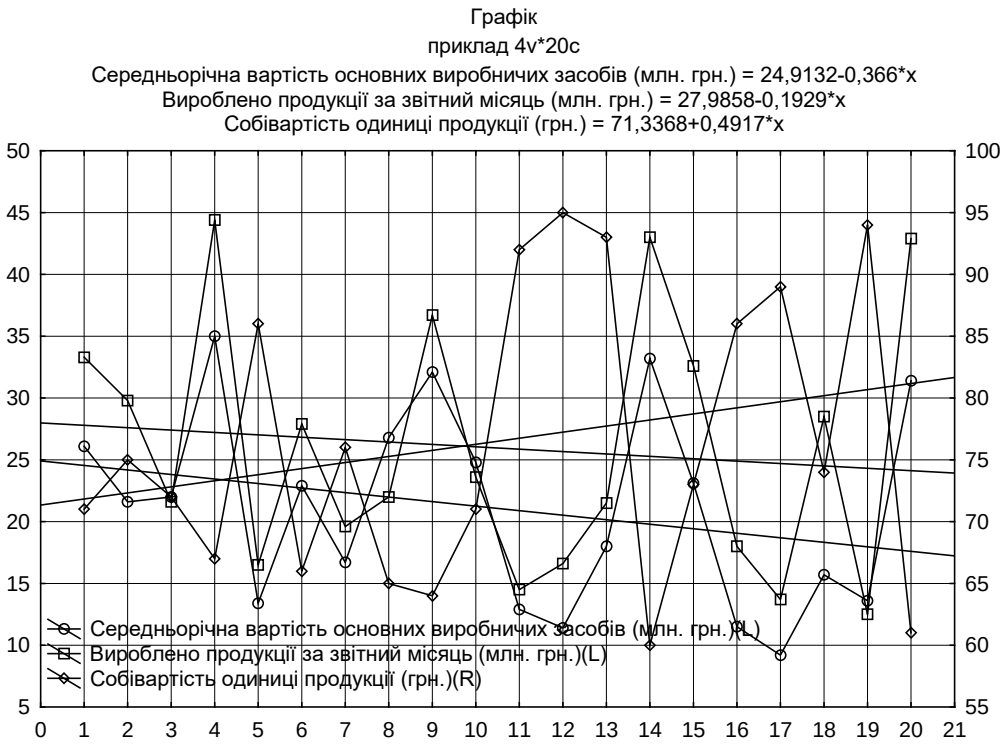


Рис. 4.14. Графік з налаштуваннями

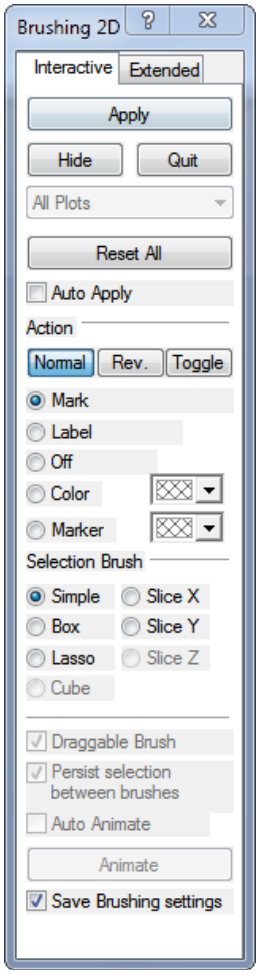


Рис. 4.15. Вікно ***Brushing 2D***

Припустимо, що значення четвертого спостереження змінної ***Вироблено продукції за звітний місяць*** є викидом. Перейдемо на вкладку ***Edit*** та в групі ***Customize Graph (Налаштування графіку)*** клацнемо по кнопці із зображенням прицілу на панелі інструментів (при підводі курсора миші до неї висвітиться напис ***Brushing***). Відкриється вікно ***Brushing 2D*** (рис. 4.15) на вкладці ***Interactive (Інтерактивне)***. У рамці ***Action (Дія)*** на вкладці ***Normal (Нормальна)*** виділимо опцію ***Label (Мітка)***. У рамці ***Selection brush (Вибір зафарбовування)*** виділимо опцію ***Simple (Простий)***. Підведемо "приціл" до точки і клацнемо лівою кнопкою миші. Якщо виділений режим ***Auto Apply (Автооновлення***), то з'явиться мітка з позначенням значення. Якщо режим ***Auto Apply*** не виділений, то натиснемо на кнопку ***Apply (Оновити)***, яка розташована у верхній частині вікна. Якщо треба виключити цю точку з графіка, на вкладці ***Normal*** виділимо опцію ***Off (Вимкнути)***, підведемо "приціл" до точки, клацнемо лівою кнопкою миші і натиснемо на кнопку ***Apply***. Точка зникне. Зверніть увагу, що після цих дій з'явиться нове рівняння регресії. Якщо провести перераховані дії, попередньо виділивши опцію ***Mark (Позначка)***, точка буде зафарбована. Можна також вибрати колір ***(Color)*** та вид ***(Marker)*** позначки. При виборі кисті ***Box (Блок)***, ***Lasso (Лассо)***, ***Slice X (Площина X)***, ***Slice Y (Площина Y)*** дії, описані раніше, можна одночасно провести з групою точок. Якщо аналізується тривимірний графік, стають активними кисті ***Cube (Куб)*** і ***Slice Z (Площина Z)***.

Якщо виділити вкладку ***Rev. (Обернена)*** (рис. 4.15), операції ***Mark***, ***Label***, ***Off*** поміняються на операції з протилежним змістом дій – ***Unmark (Виключити позначки)***, ***Unlabel (Виключити мітки)***, ***On (Включити)***. Якщо виділити вкладку ***Toggle (Переключити)***, операції ***Mark***, ***Label***, ***Off*** поміняються на ***Toggle mark (Переключити позначки)***, ***Toggle label (Переключити мітки)***, ***Toggle On/Off (Включити/виключити)*** (рис. 4.15). Кнопка ***Quit (Вихід)*** у верхній частині вікна призначена для виходу. Кнопка ***Reset All (Скасувати все)*** відмінить всі дії, що були проведені з виділеними точками.

Можна переглянути всі зміни, внесені інструментом ***Brushing.*** Для цього необхідно вибрати ***Graph Data Editor (Редагувати дані)*** з контекстного меню. У рядках, де були внесені зміни будуть зроблені позначки:

– змінна позначена;

– змінна помічена (на графіку вкажеться її значення);

– змінна вилучена.

На двовимірних графіках кожна залежність подана парою стовпців X і Y, де X – порядковий номер змінної та Y – значення змінної (рис. 4.16). Кожна пара (X, Y) відповідає точці на графіку. Тут також можна змінювати дані, видаляти точки, добавляти рядки або нову залежність; всі зроблені зміни будуть відображені на графіку.

	Line Plot		Line Plot		Line Plot	
	X	вартість основних виробничих засобів (млн.	X	продукції за звітний місяць (млн.	X	Собівартість од иниці продукції (грн.)
1	1	26,1	1	33,3	1	71
2	2	21,6	2	29,8	2	75
3	3	22	3	21,6	3	72
4	4	35	4	44,4	4	67
5	5	13,4	5	16,5	5	86
6	6	22,9	6	27,9	6	66
7	7	16,7	7	19,6	7	76
8	8	26,8	8	22	8	65
9	9	32,1	9	36,7	9	64
10	10	24,8	10	23,6	10	71
11	11	12,9	11	14,5	11	92
12	12	11,4	12	16,6	12	95
13	13	18	13	21,5	13	93
14	14	33,2	14	43	14	60
15	15	23,1	15	32,6	15	73
16	16	11,5	16	18	16	86
17	17	9,2	17	13,7	17	89
18	18	15,7	18	28,5	18	74
19	19	13,6	19	12,5	19	94
20	20	31,4	20	42,9	20	61

Рис. 4.16. Вхідні дані для побудови діаграми розсіювання

3. Тривимірні графіки

3D Graphs (Тривимірні графіки) дозволяють аналізувати дані в тривимірному просторі. ***3D Graphs*** подібні з складеними лінійними графіками, але з певними особливостями. Розглянемо основні види ***3D Sequential Graphs 3D (3D послідовні графіки)*** (рис. 4.17), що використовуються для графічного аналізу простої послідовності даних або статистичного зведення інформації групи або підмножини.

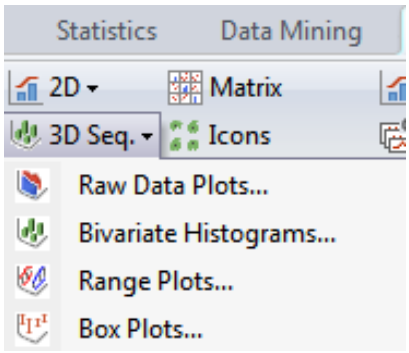


Рис. 4.17. Меню ***3D Sequential Graphs***

Raw Data Graphs (Графіки вихідних даних) показують співвідношення між значеннями змінних. Щоб побудувати графік, треба з групи ***More…*** вибрати меню ***3D Seq. (3D послідовні графіки)→Raw Data Graphs***. Відкриється діалогове вікно ***Raw Data Graphs.*** На вкладці ***Advanced*** наведено можливі типи вихідних графіків: ***Columns (Стовпці)***, ***Blocks (Блоки)***, ***Ribbon (Стрічки)***, ***Lines (Лінії)***, ***Spikes (Сплески)***, ***Contour/Discrete (Лінії рівня)***, ***Surface (Поверхня)*** та ***Contours (Контури)***.

Bivariate Histograms (Гістограми двох змінних) використовуються для візуалізації табульованих значень двох змінних або для візуалізації таблиць спряженості двох змінних.

3D Range Plots (3D діаграми діапазонів) подібні статистичним 2D діаграмам діапазонів, відображають діапазони значень або стовпці помилок, що відповідають певним точкам даних. 3D діаграми діапазонів бувають різних типів: ***Point Ranges (Точкові)***, ***Border-style Ranges (Граничні)***, ***Error Bar-style Ranges (Діапазони помилок)***, ***Double Ribbon Ranges***

(Діапазони подвійних стрічок), ***Flying Boxes (Прямокутники)*** та ***Flying Blocks (Блоки)***.

3D Box Plot (3D діаграми розмаху) подібні ***2D діаграмам розмаху***.

3D XYZ Graphs (Тривимірні графіки) – це найпростіший тип тривимірних залежностей, що використовуються для графічної інтерпретації трьох значень (X, Y, Z) в якості координат точок в 3D-просторі (рис. 4.18). До них відносяться ***Scatterplot (діаграми розсіювання)***, що відображає взаємозв'язок між трьома змінними в тривимірному просторі, при цьому кожній точці відповідає координата X, Y і Z та ***Categorized XYZ Plots (Категорізовані тривимірні графіки), які*** є одним з найпотужніших аналітичних методів дослідження.

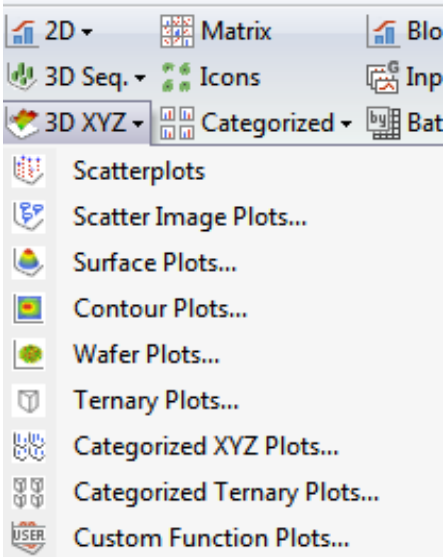


Рис. 4.18. Меню ***3D XYZ***

Щоб побудувати діаграму розсіювання, треба з групи ***More…*** вибрати меню ***3D XYZ→Scatterplot***. З'явиться діалогове вікно ***3D Scatterplot.*** Аналогічно будуються ***Категорізовані тривимірні графіки.***

Редагування тривимірних графіків аналогічні до двовимірних.

Завдання для самостійної роботи

Для даних одного файлів для самостійної роботи лабораторних робіт 2 та 3:

1) побувати двовимірну гістограму та:
   - змінити її тип;
   - задати експоненційний тип підгонки;
   - змінити тип показу на кумулятивний;
   - встановити інтервали;
   - змінити заголовок на “Завдання 1”;
2) побувати діаграму розсіювання та:
   - змінити її тип;
   - задати лінійний тип підгонки;
   - подувати еліпс нормального розподілу;
   - змінити позначення осей та масштабувати вісі,
   - змінити заголовок на “Завдання 2”;
3) побувати діаграму розмаху та:
   - змінити її тип;
   - задати поліноміальний тип підгонки;
   - вибрати центральну точку – середнє та задати середньоквадратичне відхилення для прямокутника та мінімальне-максимальне для відрізків;
   - вибрати центральну точку – медіана та задати відсотки для прямокутника та мінімальне-максимальне для відрізків;
   - змінити заголовок на “Завдання 3”;
4) побувати лінійний графік та:
   - змінити його тип на трасувальний;
   - побувати лінійний графік для багатьох змінних;
   - змінити місце розташування легенди:
   - побудувати один з графіків вдовж правої осі Y та задати її параметри та масштаб;
   - додати розриви;
   - додати лінію тренду на графік;
   - виключити викиди, помітити та позначити деякі точки;
   - показати дані графіка;
   - змінити заголовок на “Завдання 4”;
5) побувати кругову діаграму;
6) побувати всі вказані види тривимірних діаграм та здійснити підгонку до діаграми розсіювання.

Лабораторна робота № 5
Аналіз статистичних даних за допомогою модуля
Descriptive statistics (Описова статистика)

1. Основні положення модуля Descriptive statistics

Модуль ***Descriptive statistics (Описові статистики)*** об'єднує методи статистичного аналізу, що найчастіше використовуються на початковому етапі обробки даних, коли визначається структура та залежності між даними.

Для запуску модуля у вкладці ***Statistics*** у групі ***Base*** меню ***Basic Statistics*** або в меню ***Statistics*** у діалоговому вікні ***Basic Statistics and Tables*** треба вибрати ***Descriptive statistics***. Вигляд діалогового вікна ***Descriptive statistics*** зображено на рис. 5.1.

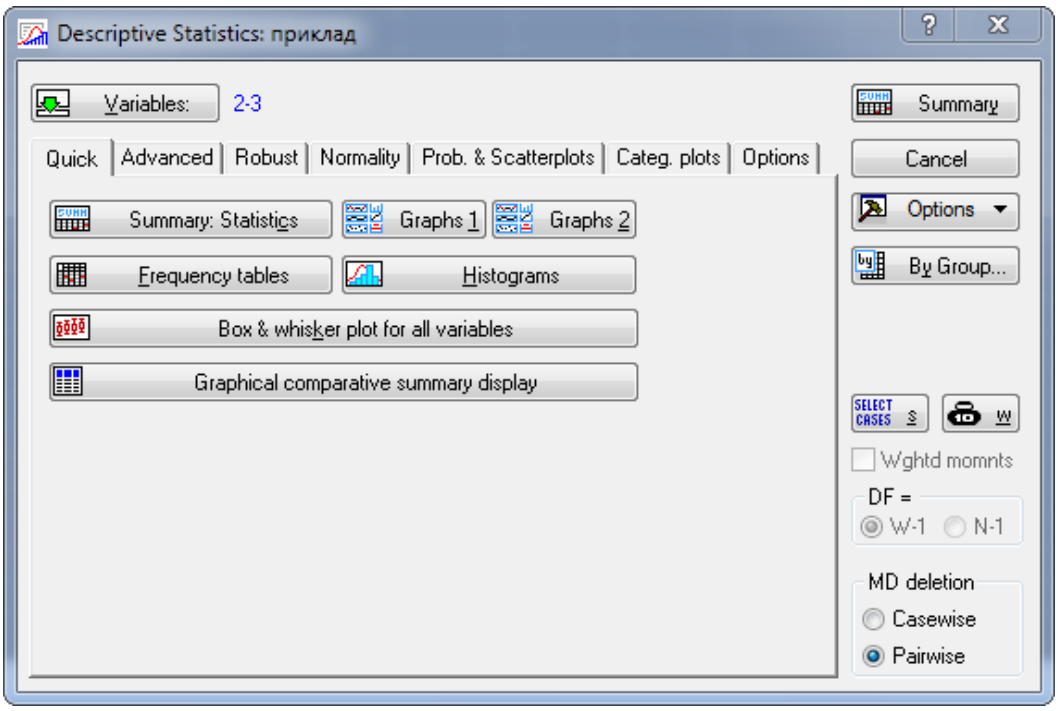


Рис. 5.1. Діалогове вікно ***Descriptive Statistics***

Для вибору змінної або діапазону змінних треба натиснути кнопку ***Variables*** і у вікні клацнути на імені змінної (змінних) (рис. 5.2). У вікні вибору змінної є доступними всі змінні файлу даних та можливі опції: ***Select All (Вибрати всі змінні)***, ***Spread (Розширений опис)*** – відображення додаткової інформації, ***Shrink*** (***Скорочений опис***) – приховування додаткової інформації, ***Zoom (Збільшити)*** **–** відображення детальної інформації про змінну (рис. 5.3).

Для перегляду результатів треба натиснути кнопку ***Summary: Statistics (Результат: Статистики)*** у діалоговому вікні ***Descriptive Statistics***. Відкриється таблиця з основними статистиками. Якщо дослідника цікавлять інші статистичні показники, їх потрібно вибрати на вкладці ***Advanced***. За допомогою кнопки ***Select all stats*** діалогового вікна ***Descriptive statistics*** вкладки ***Advanced*** можна вибрати всі статистичні показники. Для аналізу розкиду даних передбачені графіки ***Box & whisker plot for all variables*** (діаграми розмаху для всіх змінних), що доступні на вкладці ***Quick***. Крім того, на вкладці ***Quick*** розташовані кнопки ***Frequency tables***, ***Histograms***, що дозволяють провести аналіз, який описувався в попередніх лабораторних роботах.

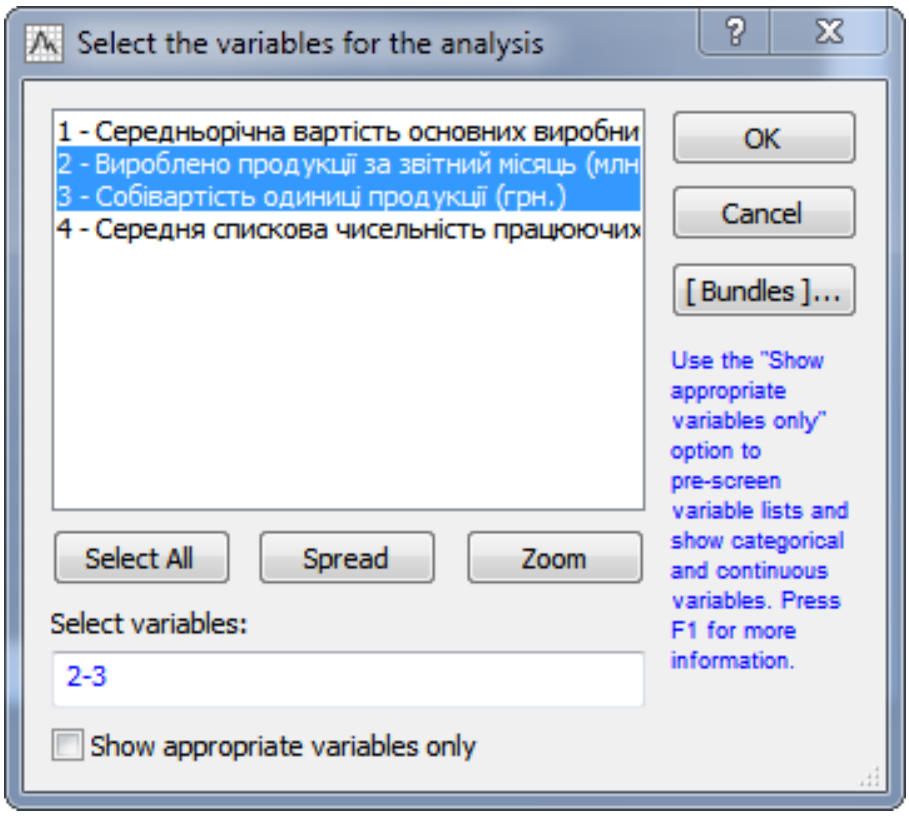


Рис. 5.2. Вибір змінної або змінних

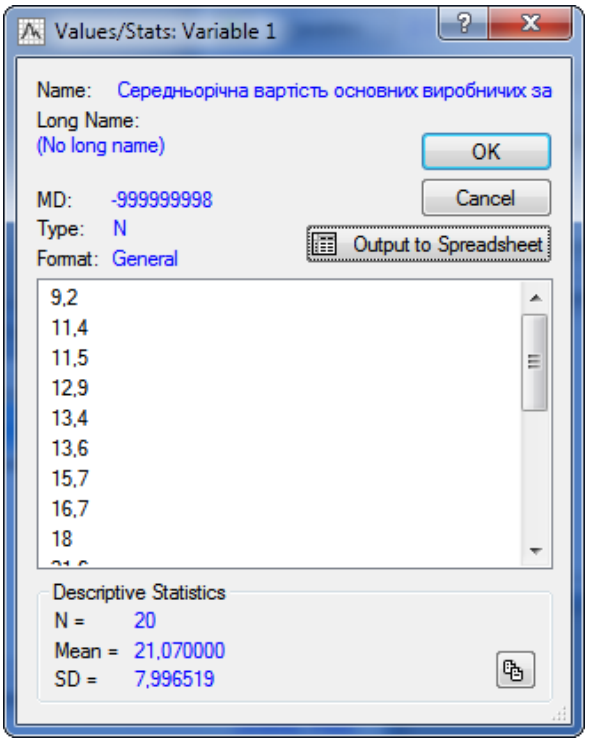


Рис. 5.3. Діалогове вікно ***Zoom***

Кнопка ***Graphs 1 (Графіки 1)*** на вкладці ***Quick*** діалогового вікна ***Descriptive statistics*** (рис. 5.1) дозволяє побудувати гістограму, графік імовірного розподілу, діаграму розмаху та подати інформацію про основні статистичні показники (рис. 5.4).

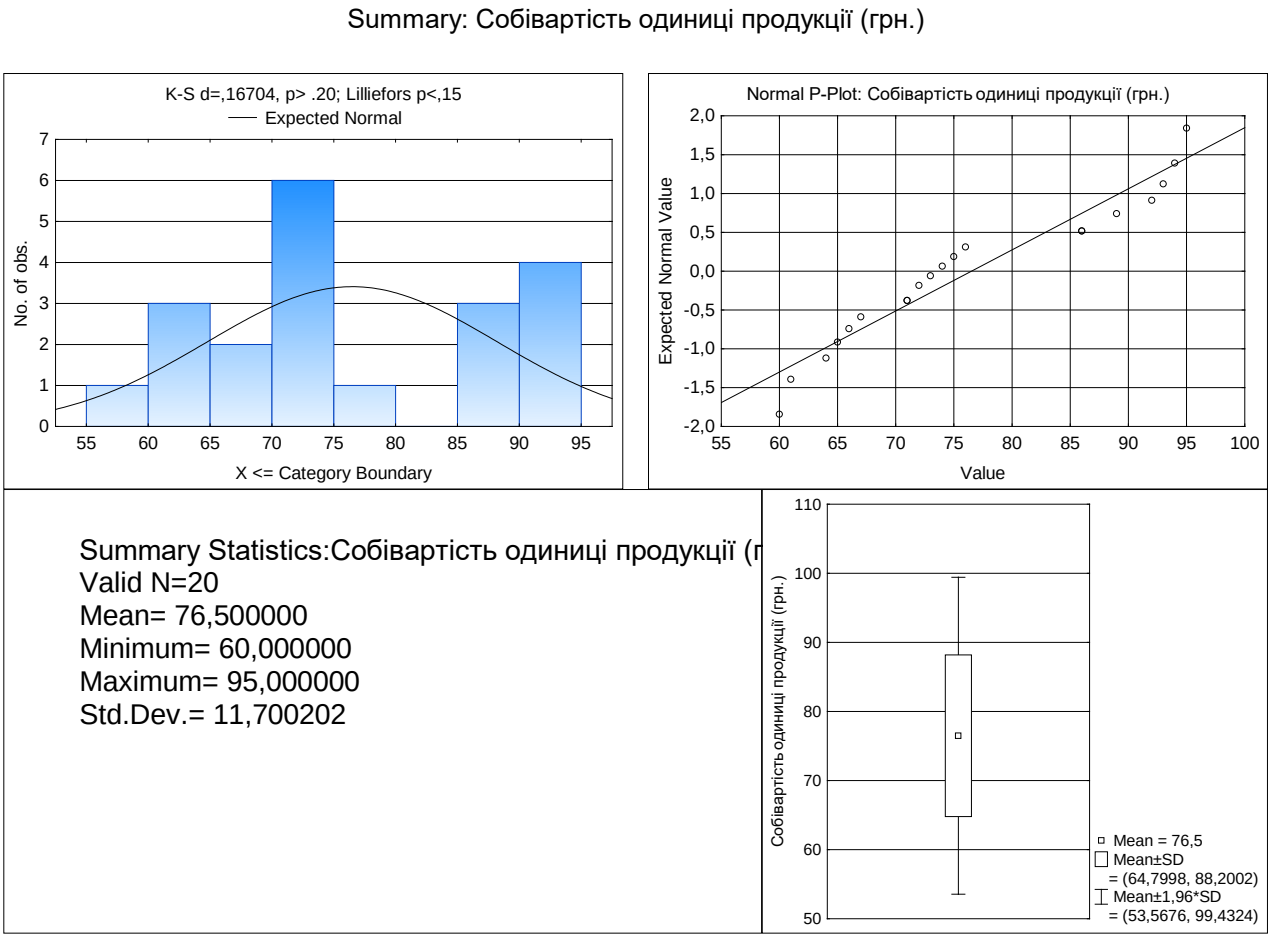


Рис. 5.4. Результат виконання ***Graphs 1*** діалогового вікна ***Descriptive statistics***

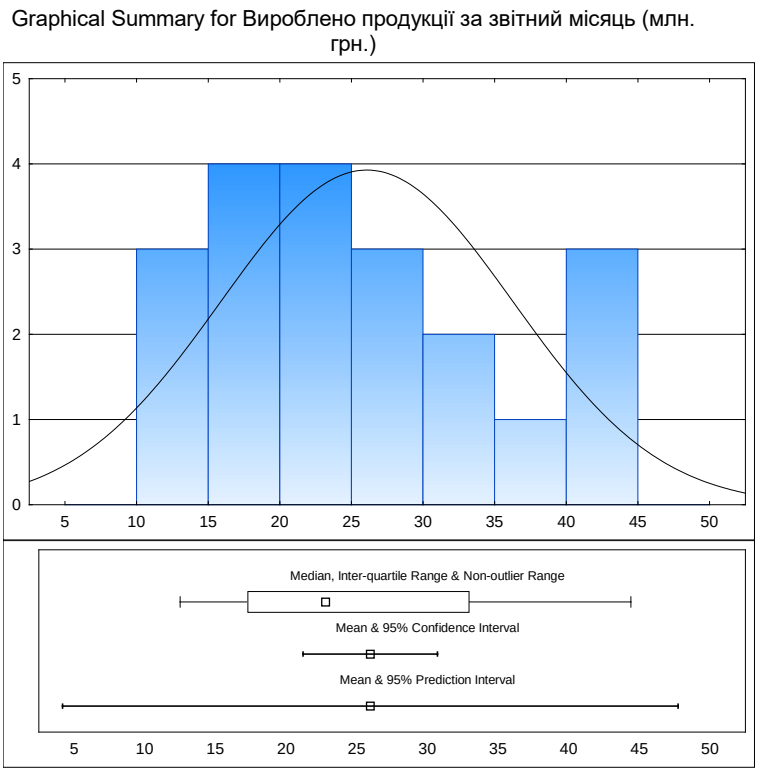
Кнопка ***Graphs 2 (Графіки 2)*** на вкладці ***Quick*** діалогового вікна ***Descriptive statistics*** дозволяє побудувати гістограму, горизонтальну діаграму розмаху та подати інформацію про основні статистичні показники. Використання ***Graphs 2*** дозволяє знайти такі ще статистичні показники, як дисперсія, асиметрія і 95% прогноз для спостережень тощо (рис. 5.5).

Кнопка ***Graphical comparative summary display (Графічне порівняння результатів)*** на вкладці ***Quick*** діалогового вікна ***Descriptive statistics*** відображає до шести змінних на одному графіку, що дозволяє легко порівнювати змінні. У робочій області подаються гістограми, діаграми розмаху та описова статистика кожної змінної (рис. 5.6).

2. Додатковий аналіз (Advanced) модуля Descriptive statistics

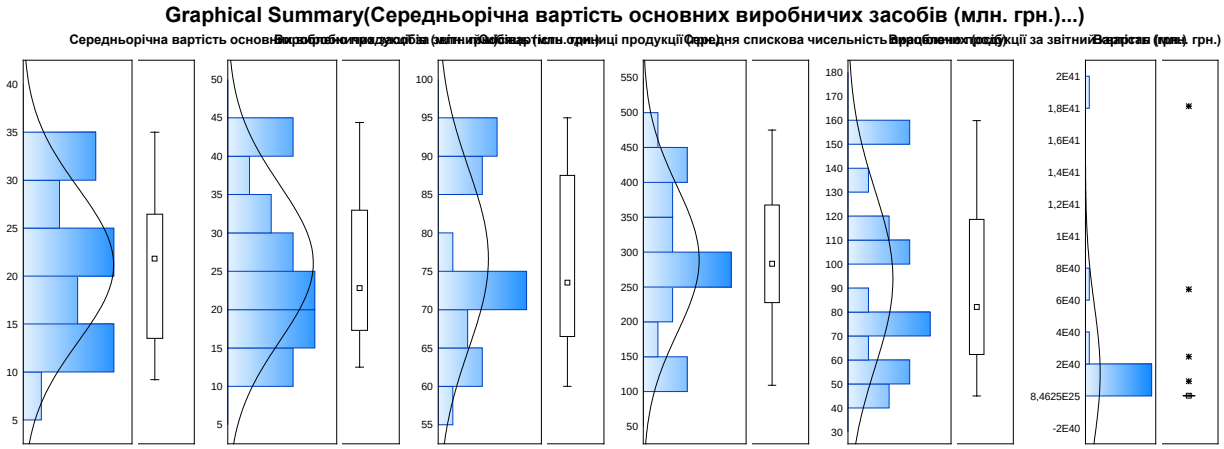
Для більш детального аналізу статистичних даних за допомогою ***Descriptive statistics*** потрібно вибрати вкладку ***Advanced*** (рис. 5.7), яка дозволяє знайти такі статистичні показники:

- ✓ ***Valid N*** – число спостережень;
- ✓ ***% valid obsvn*** – відсоток дійсних спостережень, N поділено на кількість спостережень, які пройшли критерії відбору;
- ✓ ***Mean*** – середнє значення;
- ✓ ***Sum*** – сума;
- ✓ ***Median*** – медіана;
- ✓ ***Mode*** – мода;
- ✓ ***Geom. mean*** – середня геометрична;
- ✓ ***Harm. mean*** – середня гармонійна;



Shapiro-Wilk p:	n/a
Mean:	25,96
Std.Dev.:	10,16
Variance:	103
Std.Err.Mean	2,271
Skewness:	0,537
Valid N:	20,00
Minimum:	12,50
Lower Quartile	17,30
Median:	22,80
Upper Quartile	32,95
Maximum:	44,40
95% Confidence for Std Dev	
Lower	7,723
Upper	14,83
95% Confidence for Mean	
Lower	21,21
Upper	30,71
95% Prediction for Observation	
Lower	4,180
Upper	47,74

Рис. 5.5. Результат виконання команди ***Graphs 2***



N: 20,00	N: 20,00	N: 20,00	N: 20,00	N: 20,00	N: 20,00
Mean: 21,07	Mean: 25,96	Mean: 76,50	Mean: 288	Mean: 93,46	Mean: 1,41e+040
Median: 21,80	Median: 22,80	Median: 73,50	Median: 283	Median: 82,08	Median: 9,39e+031
Min: 9,200	Min: 12,50	Min: 60,00	Min: 109	Min: 45,00	Min: 1,14e+026
Max: 35,00	Max: 44,40	Max: 95,00	Max: 475	Max: 160	Max: 1,81e+041
L-Qrt: 13,50	L-Qrt: 17,30	L-Qrt: 66,50	L-Qrt: 228	L-Qrt: 62,28	L-Qrt: 8,57e+028
U-Qrt: 26,45	U-Qrt: 32,95	U-Qrt: 87,50	U-Qrt: 368	U-Qrt: 119	U-Qrt: 2,36e+038
Variance: 63,94	Variance: 103	Variance: 137	Variance: 10979	Variance: 1337	Variance: 1,79e+081
SD: 7,997	SD: 10,16	SD: 11,70	SD: 105	SD: 36,56	SD: 4,23e+040
Std.Err: 1,788	Std.Err: 2,271	Std.Err: 2,616	Std.Err: 23,43	Std.Err: 8,175	Std.Err: 9,45e+039
Skw: 0,249	Skw: 0,537	Skw: 0,343	Skw: -0,131	Skw: 0,537	Skw: 3,686
Kurt: -1,124	Kurt: -0,851	Kurt: -1,304	Kurt: -0,575	Kurt: -0,851	Kurt: 14,25
95% Conf SD	95% Conf SD	95% Conf SD	95% Conf SD	95% Conf SD	95% Conf SD
Lower: 6,081	Lower: 7,723	Lower: 8,898	Lower: 79,68	Lower: 27,80	Lower: 3,22e+040
Upper: 11,68	Upper: 14,83	Upper: 17,09	Upper: 153	Upper: 53,40	Upper: 6,17e+040
95% Conf Mean	95% Conf Mean	95% Conf Mean	95% Conf Mean	95% Conf Mean	95% Conf Mean
Lower: 17,33	Lower: 21,21	Lower: 71,02	Lower: 239	Lower: 76,35	Lower: -5,70e+039
Upper: 24,81	Upper: 30,71	Upper: 81,98	Upper: 337	Upper: 111	Upper: 3,39e+040

Рис. 5.6. Графічне порівняння результатів

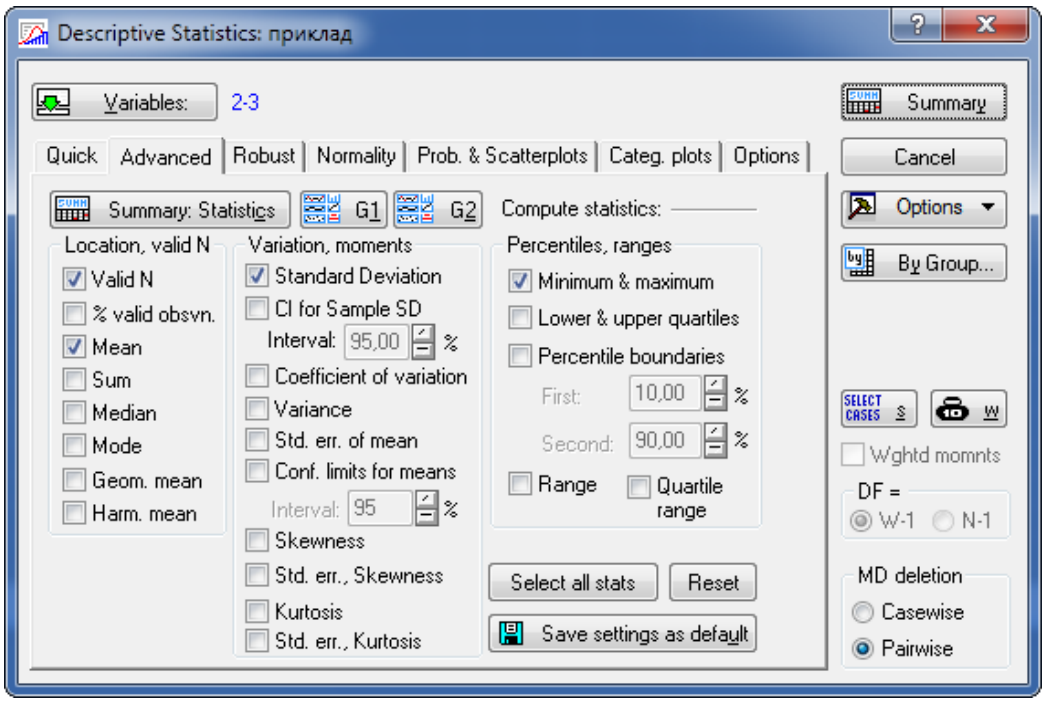


Рис. 5.7. Вкладка ***Advanced*** діалогового вікна ***Descriptive statistics***

- ✓ ***Standard Deviation*** – середньоквадратичне відхилення середнього значення;
- ✓ ***CI for Sample SD*** – довірчий інтервал для середньоквадратичного відхилення;
- ✓ ***Coefficient of Variation*** – коефіцієнт варіації;
- ✓ ***Variance*** – дисперсія;
- ✓ ***Std. err. of mean*** – стандартна помилка середнього значення;
- ✓ ***Conf. limits of mean*** – межі достовірності середнього значення;
- ✓ ***Skewness*** – коефіцієнт асиметрії;
- ✓ ***Std. err., Skewness*** – стандартна помилка значень асиметрії;
- ✓ ***Kurtosis*** – коефіцієнт ексцесу;
- ✓ ***Std. err., Kurtosis*** – стандартна помилка значень ексцесу;
- ✓ ***Minimum & maximum*** – мінімум і максимум;
- ✓ ***Lower & upper quartiles*** – нижній та верхній квартилі (тобто межі, що відтинають нижні та верхні 25% значень вибірки);
- ✓ ***Percentile boundaries*** – межі персентилів;
- ✓ ***Range*** – розмах;
- ✓ ***Quartile range*** – амплітуда варіювання квартилів.

Команди на вкладці ***Robust (Стійкі)*** діалогового вікна ***Descriptive statistics*** дозволяють розраховувати надійні значення, які не чутливі до викидів.

У вкладці ***Normality (Відповідність нормальному розподілу)*** діалогового вікна ***Descriptive statistics*** можна обчислити такі критерії:

- ***Kolmogorov-Smirnov test (критерій Колмогорова-Смірнова)*** використовується при відомому середньому і середньоквадратичному відхиленні генеральної сукупності. Якщо обчислена D-статистика значуща, то гіпотеза про те, що дані мають нормальний розподіл, відкидається.
- ***Lilliefors test (критерій Ліллієфорса)*** використовується при невідомому середньому і середньоквадратичному відхиленні генеральної сукупності (оцінюються за наявними даними). Якщо обчислена D-статистика значуща, то гіпотеза про те, що дані мають нормальний розподіл, відкидається.

- ***Shapiro-Wilk's W test (критерій Шапіро-Уїлка W)*** за даними спостережень обчислюється W-статистика, і якщо вона значуща, гіпотеза про нормальний характер розподілу відкидається, інакше – приймається.

Обчислені критерії будуть подані в таблиці частот. У цілому, якщо рядки у таблиці для якогось ряду даних (змінної) виділені червоним кольором, то це говорить про необхідність відкинути гіпотезу про нормальний характер розподілу даної змінної.

Вказані критерії обчислюються, якщо задані параметри ***Categorization*** вкладки ***Normality***:

- ***Number of intervals*** – кількість інтервалів;
- ***Integer intervals (categories)*** – цілі інтервали (категорії) (наприклад, стать, код населеного пункту тощо). Якщо вибрати цілі інтервали, то вказані критерії стануть неактивними.

Вкладка ***Options*** діалогового вікна ***Descriptive statistics*** дозволяє знайти додаткові опції. Зокрема, можна обрати ***Options for descriptive statistics (Опції для описової статистики)*** та ***Options for Box-Whisker plots (Опції для діаграм розмаху)***:

– для описової статистики:
  - ✓ ***Display long variable names*** – довгі текстові назви змінних;
  - ✓ ***Extended precision calculations*** – подвійна точність розрахунків (до 10 значущих цифр);

– для діаграми розмаху:
  - ✓ ***Median/Quartiles/Range*** – показ в центрі діаграми медіани, квартилів прямокутниками та розмаху відрізками (***Медіана/квартиль/розмах***);
  - ✓ ***Mean/SE/SD*** – середнє/стандартна похибка/середньоквадратичне відхилення;
  - ✓ ***Mean/SD/1.96*SD*** – середнє/середньоквадратичне відхилення/ 1,96*середньоквадратичне відхилення;
  - ✓ ***Mean/SE/1.96*SE*** – середнє/стандартна похибка/1,96*стандартна похибка.

У рамці ***MD deletion (Видалення пропущених даних)*** вкладки ***Options*** можна задати режими роботи з пропущеними даними:

✓ якщо вибрано ***Casewise*** – STATICTICA проігнорує всі спостереження (рядки), де пропущені дані;

✓ якщо вибрано ***Pairwise*** – всі дані будуть включені в статистичний аналіз.

Інші вкладки у модулі ***Descriptive statistics*** призначені для побудови статистичних графіків. Вкладка ***Prob. & Scatterplots (Графіки ймовірності та діаграми розсіювання)*** надає широкий набір способів графічного дослідження змінних. Вкладка ***Categ.plots (Категоризовані графіки)*** призначена для графічного дослідження змінної з попереднім проведенням категоризації (розбиття на різні підмножини).

3. Типовий приклад

Відомий розподіл студентів (табл. 5.1) за віком на різних факультетах університету. Визначити по кожному з факультетів основні статистичні показники (наприклад, середній вік студентів, моду, медіану тощо). Побудувати діаграми розмаху для середнього значення та медіани. Порівняйте дані економічного та юридичного факультетів.

Таблиця 5.1

Вік студента (роки)	*Кількість студентів на факультетах університету (чол.)*			
	економічний	*іноземних мов*	*філологічний*	*юридичний*
18-20	*250*	*220*	*150*	*260*
20-28	*1000*	*950*	*700*	*1100*
28-30	*600*	*450*	*340*	*500*

Вік студента (роки)	Кількість студентів на факультетах університету (чол.)			
	економічний	іноземних мов	філологічний	юридичний
30-34	120	100	80	130
34-40	50	30	20	40
старше 40	10	2	5	12

Розв'язування. *Для знаходження основних статистичних показників використаємо модуль описової статистики. Для визначення статистик на вкладці* ***Advanced*** *виберемо необхідні статистики (рис. 5.8). Результати будуть подані в таблиці (рис. 5.9). Для побудови діаграм розмаху спочатку потрібно задати її центр на вкладці* ***Options*** *(виберемо або* ***Median/Quartiles/Range*** *(5.10 а), або* ***Mean/SE/SD*** *(5.11 а)) . Вибравши вкладку* ***Quick*** *,натиснемо кнопку* ***Box & whisker plot for all variables****. Діаграми розмаху показані на рис. 5.10 б, 5.11 б. Як видно з графіків, на економічному та юридичному факультетах вікова структура студентів майже однакова. Детальніше порівняти вікову структуру студентів економічного та юридичного факультетів можна також використовуючи вкладку* ***Quick →Graphical comparative summary display*** *(при цьому вибравши спочатку відповідні змінні, тобто економічний та юридичний факультети). Результат порівняння вікової структури студентів обох факультетів показаний на рис. 5.12.*

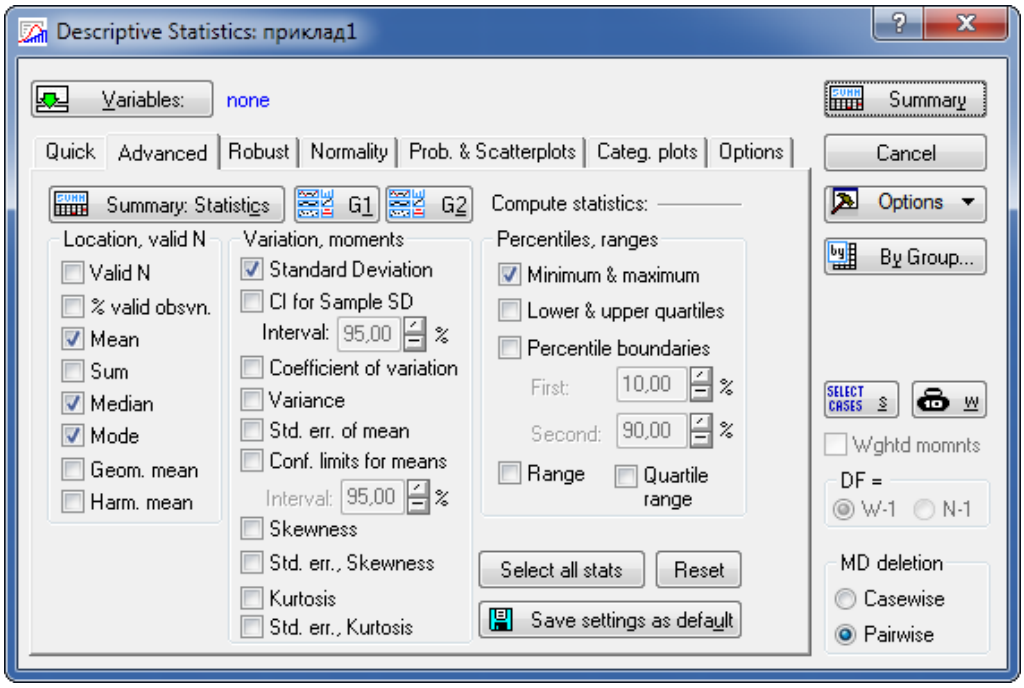
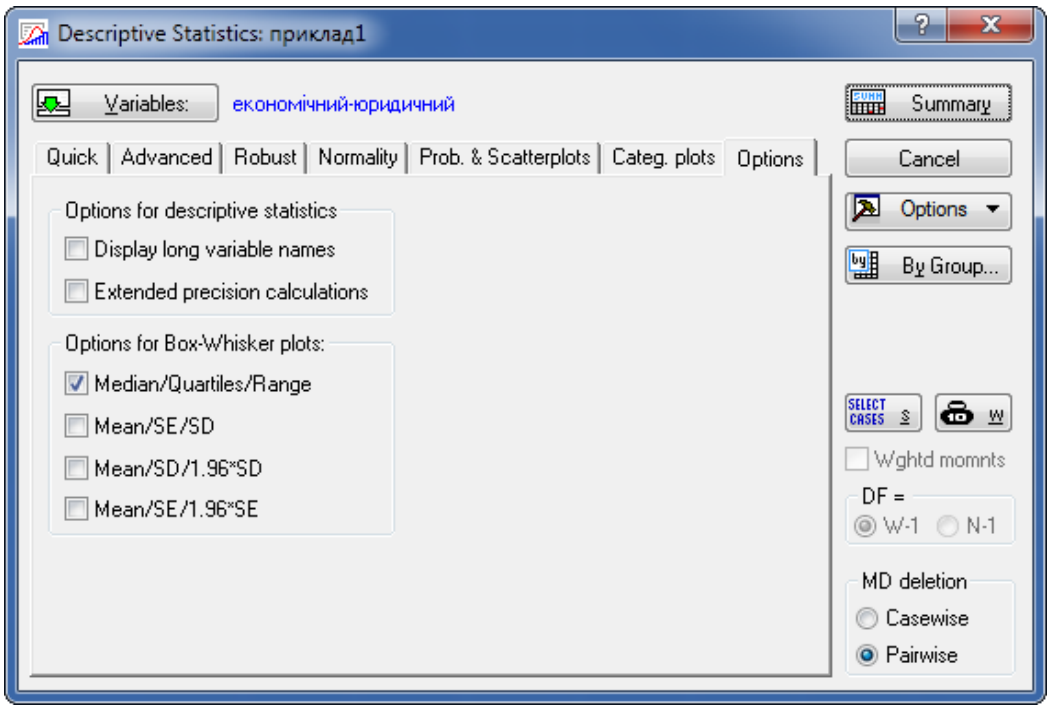


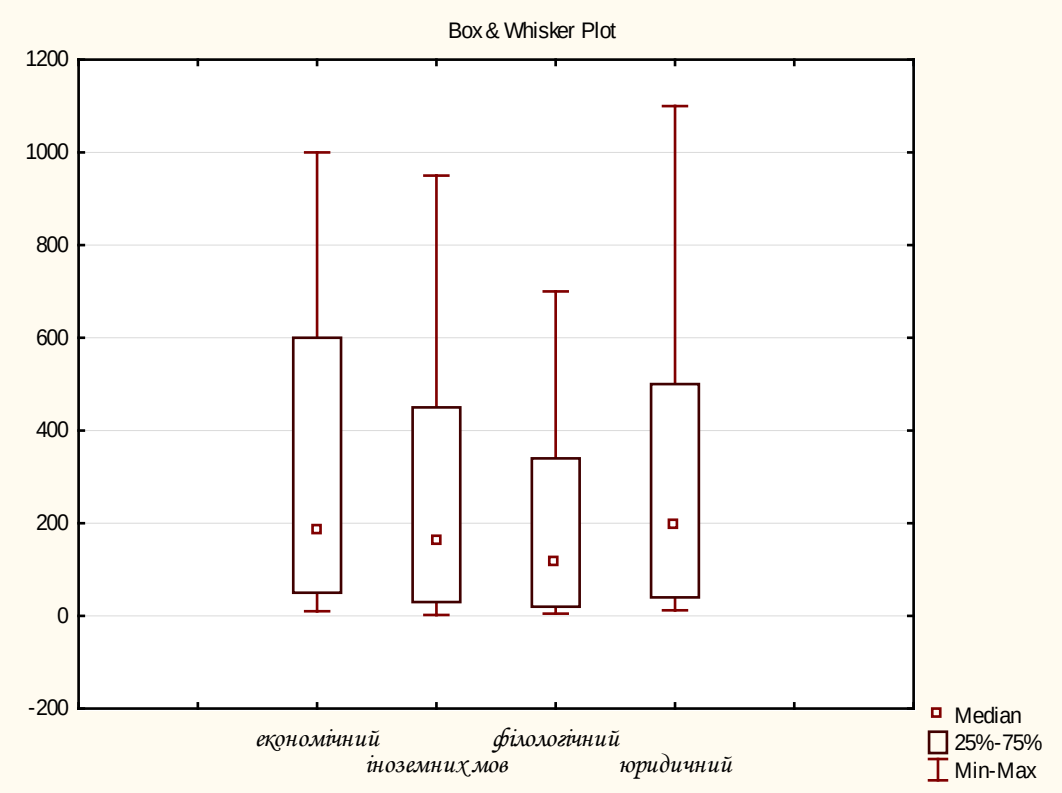
Рис. 5.8. Вибір основних статистичних показників вкладці ***Advanced*** *діалогового вікна* ***Descriptive statistics***

Variable	Descriptive Statistics (приклад1)						
	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Std.Dev.
економічний	338,3333	185,0000	Multiple	1	10,00000	1000,000	387,9905
іноземних мов	292,0000	160,0000	Multiple	1	2,00000	950,000	361,3087
філологічний	215,8333	115,0000	Multiple	1	5,00000	700,000	266,6161
юридичний	340,3333	195,0000	Multiple	1	12,00000	1100,000	412,6217

Рис. 5.9. Таблиця результатів модуля ***Descriptive statistics***

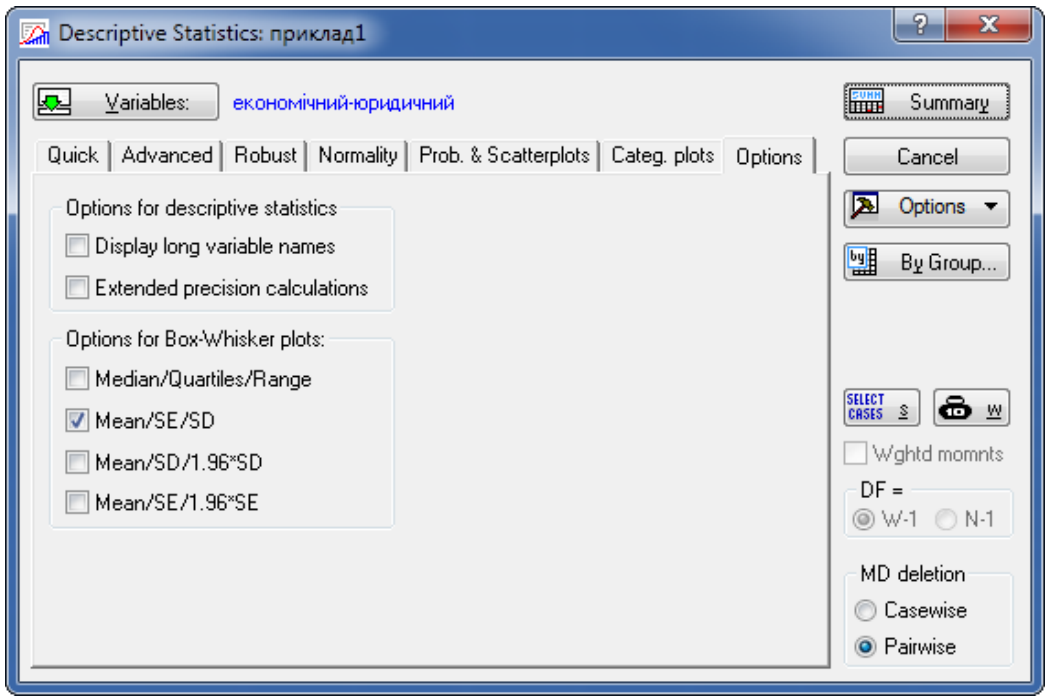


а)

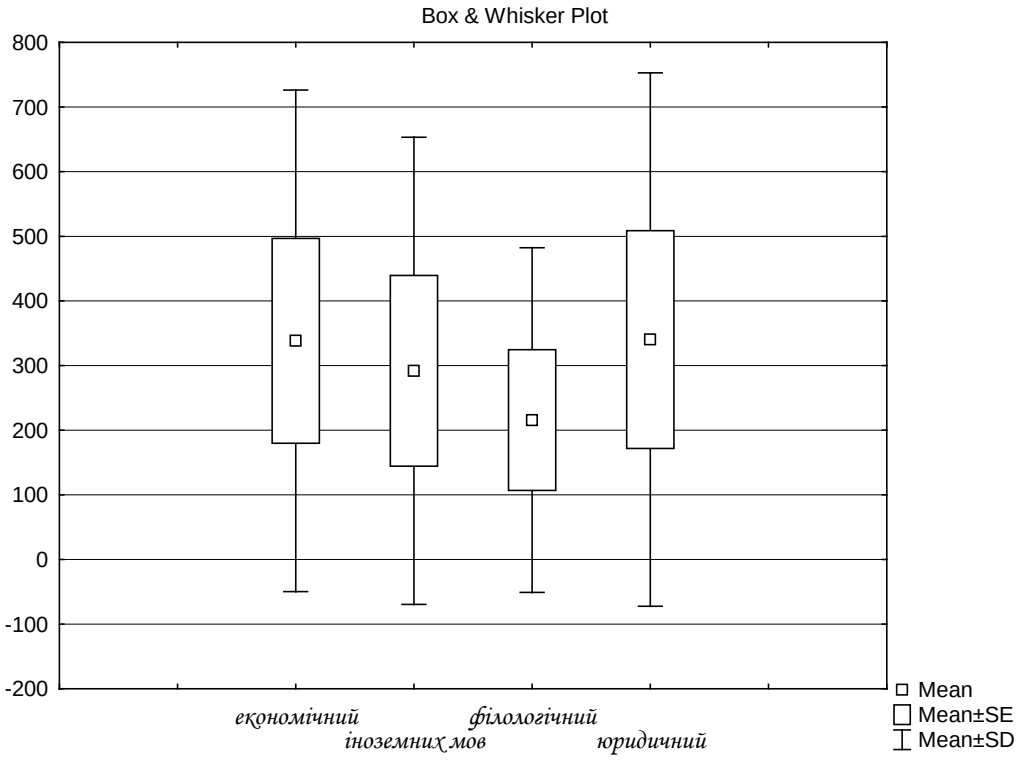


б)

*Рис. 5.10. Вибір центру на діаграмі (вкладка **Options**) та діаграма розмаху для центру медіани*



а)



б)

Рис. 5.11. Вибір центру на діаграмі (вкладка ***Options****) та діаграма розмаху для центру середнє значення*

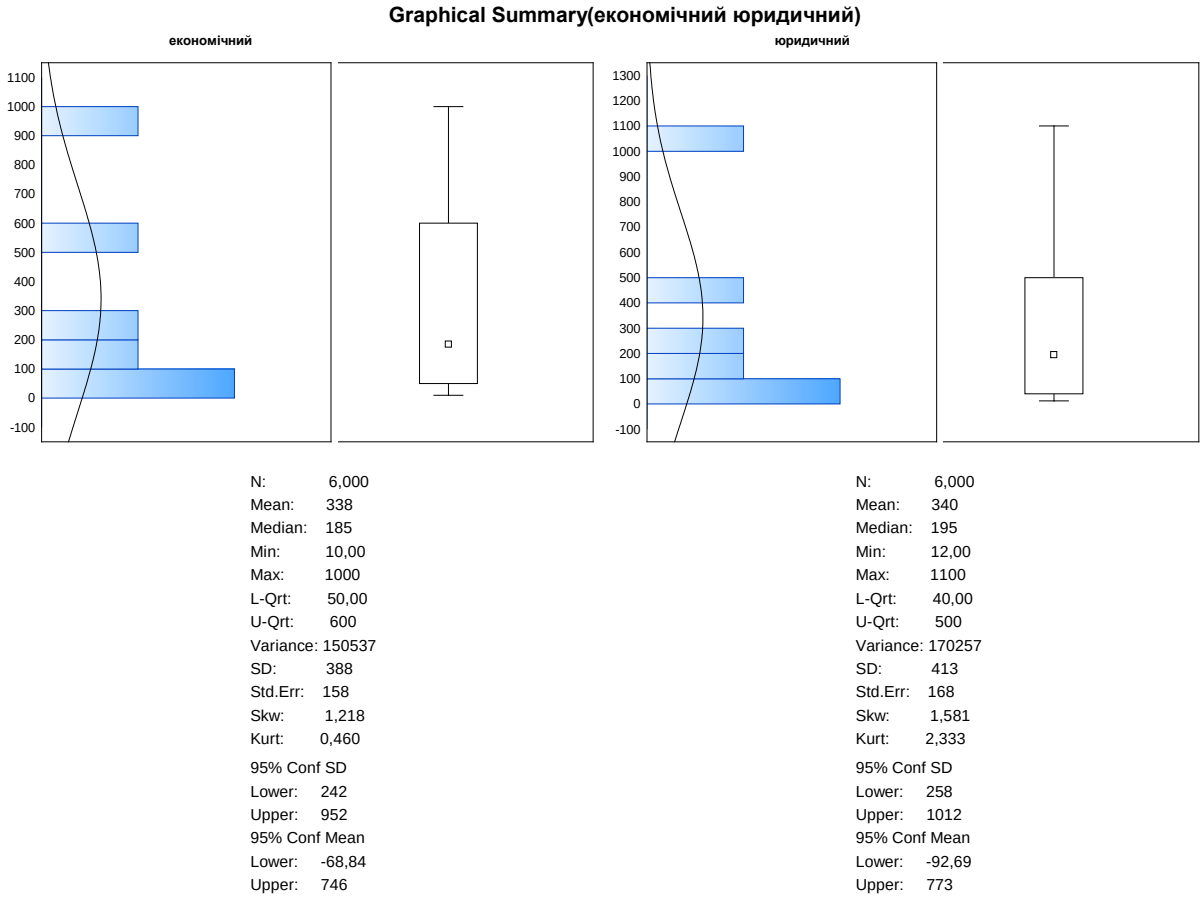


Рис. 5.12. Результат порівняння вікової структури по факультетах

Завдання для самостійної роботи

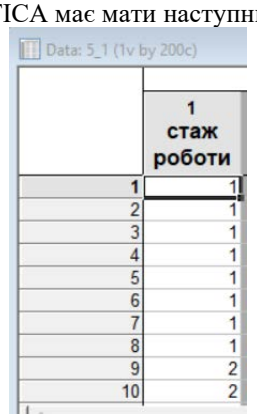
5.1. У таблиці 5.2 подано дані про стаж роботи 200 робітників деякого підприємства.

Таблиця 5.2

1	2	3	4	4	8	8	3	1	3	9	7	4	4	4	9	9	7	5	5	7	7	7	6	6
1	2	3	3	4	8	4	8	9	3	5	4	5	5	5	5	5	5	7	10	7	7	6	6	7
1	3	4	7	3	9	4	4	5	3	5	4	4	9	5	5	9	8	10	5	5	7	6	6	6
1	2	4	7	3	5	5	9	3	3	5	5	4	4	5	5	7	10	7	5	7	6	6	6	6
1	2	4	5	3	8	5	9	3	8	5	5	4	5	5	5	10	7	7	5	6	6	6	6	6
1	2	4	3	8	9	2	5	3	5	5	4	4	10	7	7	7	7	5	5	6	6	6	6	6
1	2	3	3	4	5	8	4	5	6	5	5	5	4	5	8	8	8	5	4	8	8	6	6	5
1	2	2	3	2	2	8	5	5	5	8	7	4	4	7	7	7	5	5	6	6	6	6	6	6

Обчислити основні статистичні показники, використовуючи модуль ***Descriptive statistics***: середню, довірчий інтервал для середньої, медіану, мінімум, максимум, нижній та верхній квартилі, розмах варіації, квартальний розмах, дисперсію, середнє квадратичне відхилення, стандартну помилку для середнього квадратичного відхилення, коефіцієнти асиметрії та ексцесу. Пояснити значення всіх статистичних показників і сформувати відповідні висновки.

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх 200 спостережень.



5.2. За офіційними статистичними даними про рівень зайнятості населення за статтю, віковими групами та місцем проживання у 2011 році, що розміщенні на сайті Державної служби статистики України[1] здійснити аналіз статистичних даних, використовуючи модуль ***Descriptive statistics***.

Обчислити основні статистичні характеристики; за згрупованими даними побудувати гістограми і діаграми розмаху; сформувати висновок про близькість отриманого розподілу до нормального. На основі аналізу показників форми розподілу і за допомогою критеріїв сформувати висновки.

5.3. У таблиці 5.3 подано дані вибіркового спостереження про товарообіг за місяць типових торгових організацій двох мікрорайонів міста (обстежувалося 20 підприємств кожного мікрорайону).

Таблиця 5.3

Товарообіг підприємств першого мікрорайону (тис. грн.)	*Товарообіг підприємств другого мікрорайону (тис. грн.)*
8609,2	8490,7
9139,6	8829,5
9378,4	9168,3
9526,8	9216,7

[1] http://www.ukrstat.gov.ua/

Товарообіг підприємств першого мікрорайону (тис. грн.)	*Товарообіг підприємств другого мікрорайону (тис. грн.)*
9417,3	9361,9
9091,2	8781,1
8448,1	8442,3
9623,6	9313,5
9236,4	8926,3
9042,8	8732,7
9188,0	8877,9
9430,0	9119,9
9023,6	8974,7
8994,4	8684,3
8946,0	8635,9
9333,2	9023,1
9381,6	9071,5
9275,2	9265,1
8897,6	8587,5
8849,2	8539,1

За допомогою модуля ***Descriptive statistics*** здійснити порівняння статистичних даних і сформувати відповідні висновки.

5.4. Відомі дані про величину денного виробітку водіїв міських автобусів двох автотранспортних підприємств (АТП) (табл. 5.4). Здійснити порівняння статистичних даних, обчислити основні статистичні показники, побудувати таблиці частот та гістограми. Проаналізувати отримані результати.

Таблиця 5.4

Величина виробітку	*АТП №*	*Величина виробітку*	*АТП №*	*Величина виробітку*	*АТП №*	*Величина виробітку*	*АТП №*
8274,1	1	9392,9	2	8786,3	2	8399,7	1
8516,7	2	9460,3	2	8853,7	2	8462,5	1
8584,1	2	9467,3	1	8902,1	1	8651,5	2
8525,3	1	8449,3	2	9527,7	2	8718,9	2
8588,1	1	8713,7	1	9404,5	1	9341,7	1
8650,9	1	8776,5	1	8200,4	2	9662,5	2
8921,1	2	8839,3	1	9153,3	1	9729,9	2
8988,5	2	9595,1	2	9123,3	2	8964,9	1
9055,9	2	9216,1	1	9190,7	2	9090,5	1
8336,9	1	9278,9	1	9027,7	1	9325,5	2

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх 40 спостережень.


Data: 5_3 (2v by 40c)

	1 Величина виробітку	2 АТП №
1	8274,1	1
2	8516,7	2
3	8584,1	2
4	8525,3	1
5	8588,1	1
6	8650,9	1
7	8921,1	2
8	8988,5	2
9	9055,9	2
10	8336,9	1

5.5. З метою підвищення ефективності продажів дві фірми скористалися можливостями рекламної кампанії. Є дані про щоденний товарообіг кожної фірми за 20 днів до проведення рекламної акції і за 20 днів після того (табл. 5.5).

Визначити по кожній фірмі, чи виявилася рекламна акція ефективною та знайти основні статистичні показники.

Таблиця 5.5

Дані по фірмі А		*Дані по фірмі В*	
Щоденний товарообіг до проведення рекламної акції	*Щоденний товарообіг після проведення рекламної акції*	*Щоденний товарообіг до проведення рекламної акції*	*Щоденний товарообіг після проведення рекламної акції*
8442,3	8619,0	8684,3	8794,4
8490,7	8647,3	8732,7	8842,8
8539,1	8675,6	8781,1	8891,2
8587,5	8703,9	8829,5	8939,6
8635,9	8732,2	8877,9	8988,0
8684,3	8760,5	8926,3	8400,7
9071,5	9047,2	8974,7	9084,8
9119,9	9095,6	9023,1	9133,2
9168,3	9144,0	9071,5	9181,6
9216,7	9192,4	9119,9	9230,0
9265,1	9240,8	9168,3	9278,4
9313,5	9289,2	9216,7	9326,8
9361,9	9337,6	9265,1	9375,2
8732,7	8788,8	9313,5	9423,6
8781,1	8817,1	9361,9	9472,0
8829,5	8845,4	8442,3	8552,4
8877,9	8873,7	8490,7	8600,8
8926,3	8902,0	8539,1	8649,2
8974,7	8950,4	8587,5	8697,6
9023,1	8998,8	8635,9	8746,0

Лабораторна робота № 6
Дисперсійний аналіз

1. Основні теоретичні відомості проведення дисперсійного аналізу в STATISTICA

Порівняння середніх є одним із способів виявлення залежностей між змінними. Так, наприклад, якщо при розбитті об'єктів дослідження на підгрупи за допомогою категоріальної незалежної змінної (предиктора) вірна гіпотеза про нерівність середніх деякої залежної змінної в підгрупах, то це означає, що існує імовірнісний зв'язок між цією залежною змінною і категоріальним предиктором. Найбільш загальним методом порівняння середніх є дисперсійний аналіз – ***ANOVA (Analysis of Variance)***. У термінології дисперсійного аналізу категоріальний предиктор називається фактором.

Дисперсійний аналіз можна визначити як статистичний метод, призначений для оцінки впливу різних чинників на результат експерименту, а також для подальшого планування експериментів. За допомогою дисперсійного аналізу можна досліджувати залежність кількісної ознаки (залежної змінної) від одного або декількох якісних ознак (чинників).

Дисперсійний аналіз в системі **STATISTICA** можна здійснити вибравши у вкладці ***Statistics*** у групі ***Base*** або в меню ***Statistics*** команду ***ANOVA***.

З'явиться діалогове вікно ***General ANOVA/MANOVA (Загальний дисперсійний/багатофакторний аналіз)*** (рис. 6.1).

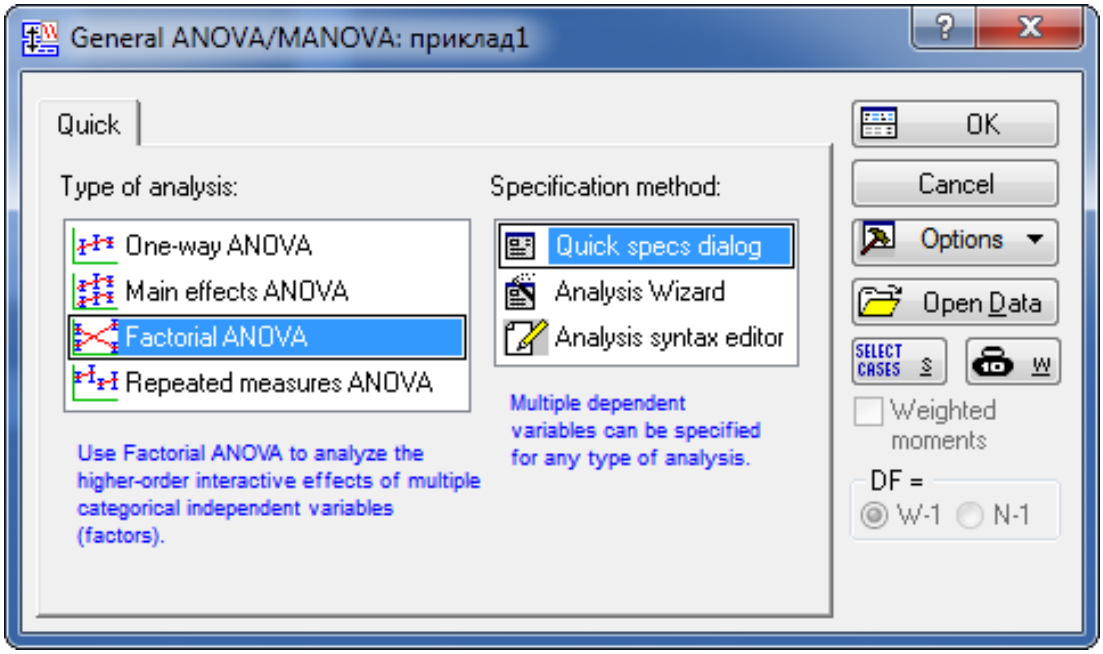


Рис. 6.1. Діалогове вікно ***General ANOVA/MANOVA***

Діалогове вікно ***General ANOVA/MANOVA*** містить два списки ***Type of analysis (Тип дисперсійного аналізу)*** і ***Specification method (Специфікація методу)***.

Список ***Type of analysis*** складається з чотирьох типів дисперсійного аналізу:

- ***One-way ANOVA (Однофакторний дисперсійний аналіз)***;
- ***Main effects ANOVA (Дисперсійний аналіз головних компонент)***;
- ***Factorial ANOVA (Багатофакторний дисперсійний аналіз)***;
- ***Repeat measures ANOVA (Дисперсійний аналіз повторних вимірювань)***.

Список ***Specification method*** дозволяє задати три типи інтерфейсу дисперсійного аналізу в **STATISTICA**:

- ***Quick Specs Dialog (Діалог швидких специфікацій)***;
- ***Analysis Wizard (Майстер аналізу)***;
- ***Analysis syntax editor (Редактор коду)***.

У діалозі ***Quick Specs Dialog*** можна задати залежні змінні і категоріальні змінні (предиктори). Різна кількість і тип змінних залежить від вибраного виду аналізу в списку ***Type of analysis*** (рис. 6.1).

Діалог ***Analysis Wizard*** призначений для задання аналізу по кроках в межах вибраної моделі. У кінці аналізу можна обчислити результати або використовувати ***Analysis syntax editor*** для подальшого налаштування за допомогою вбудованих команд, відкрити існуючий файл з командами або зберегти його для подальшого використання.

Діалог ***Analysis syntax editor*** дозволяє повністю налаштувати як параметри плану, так і параметри обчислювальних процедур. У разі потреби можна зберегти файл з готовим кодом аналізу для подальшого використання або відкрити той, що вже існує.

Після вибору методу ***Specification method*** задають тип дисперсійного аналізу ***Type of analysis***.

One-way ANOVA дозволяє оцінити вплив однієї групувальної змінної (одного міжгрупового фактора) на одну або більше залежних змінних.

Для аналізу ***Main effects ANOVA*** в діалозі ***Quick Specs Dialog*** можна задати до чотирьох категоріальних предикторів. Потім програма проведе оцінку моделі головних компонент. Даний метод дисперсійного аналізу часто використовується в аналізі для оцінки впливу великої кількості факторів, а також при аналізі збалансованих неповних наборів даних.

На відміну від розглянутих методів дисперсійного аналізу в ***Factorial ANOVA*** враховується ще одне можливе джерело мінливості – взаємодія факторів. Дані містять змінні, що є комбінаціями різних рівнів двох або більше категоріальних предикторів. Зокрема, повні факторні дані – всі можливі комбінації рівнів категоріальних предикторів. У діалозі ***Quick Specs Dialog*** також можна задати до чотирьох категоріальних предикторів.

У ***Repeat measures ANOVA*** залежні змінні містять значення одного фактора повторних вимірювань. У діалозі ***Quick Specs Dialog*** також можна задати до чотирьох категоріальних предикторів і дві або більш залежні змінні, які будуть проінтерпретовані програмою як повторні вимірювання одного фактора.

Для всіх типів дисперсійного аналізу, якщо необхідно використовувати п’ять або більш категоріальних предикторів, потрібно вибрати модуль ***GLM (Загальні лінійні моделі)***.

Етапи проведення всіх типів дисперсійного аналізу однакові, розглянемо тільки деякі з них.

Для того, щоб задати план багатофакторного дисперсійного аналізу, виберіть ***Factorial ANOVA*** як вид аналізу і ***Quick Spec Dialog*** в списку ***Specification method*** на вкладці ***Quick*** стартової панелі дисперсійного аналізу. Відкриється діалогове вікно ***ANOVA/MANOVA Factorial ANOVA***. Щоб вибрати змінні, необхідно на вкладці ***Quick*** натиснути кнопку ***Variables***. У вікні, що з’явилося, вибираємо залежні та категоріальні змінні. Якщо число залежних змінних – більше 1, то програма здійснить багатовимірний дисперсійний аналіз. Система STATISTICA підказує придатні змінні для проведення аналізу. Для цього необхідно вибрати ***Show appropriate variable only (Показувати тільки відповідні змінні)*** (рис. 6.2).

Необов’язково коди задавати вручну, оскільки програма задасть за умовчанням всі коди вибраних змінних. Однак, якщо необхідно вручну задати коди для міжгрупових чинників, натисніть кнопку ***Factor Codes (Коди чинників)***.

На вкладці ***Options*** можна:

- задати тип параметризації моделі;
- вказати тип суми квадратів (SS);
- включити крос-перевірку.

2. Методи подання результатів аналізу

Діалогове вікно (рис. 6.3) ***ANOVA Results 1 (Результати аналізу)*** з набором вкладок, яке з'являється після вибору змінних, дозволяє всесторонньо відобразити результати аналізу у вигляді таблиць і графіків.

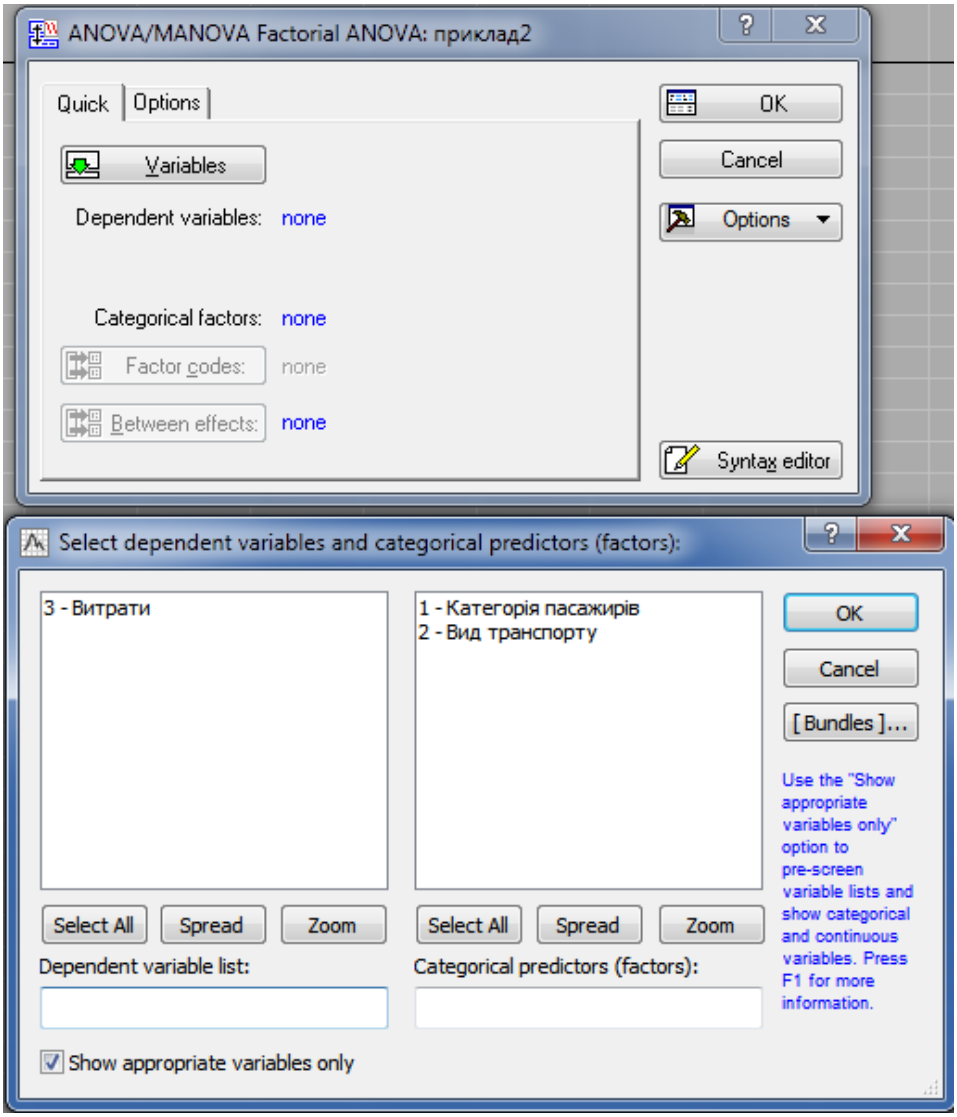


Рис. 6.2. Вибір змінних для ***Factorial ANOVA***

Вкладка ***Quick*** призначена для доступу до відображення основних результатів аналізу. ***All effects/Graphs (Всі ефекти/графіки)*** відкриває ***Table of All Effects*** (Таблицю всіх ефектів). ***All effects (Всі ефекти)*** подає результати аналізу у вигляді таблиці. ***Effect sizes (Ефект розмірів)*** дозволяє відобразити таблицю результатів з врахуванням впливу розмірів і повноважень. У полі ***Alpha values (Значення альфа)*** можна задати ***Confidence limits (Довірчий інтервал)*** та ***Significance level (Рівень значущості)***.

Розглянемо команду ***All effects/Graphs*** (рис. 6.3).

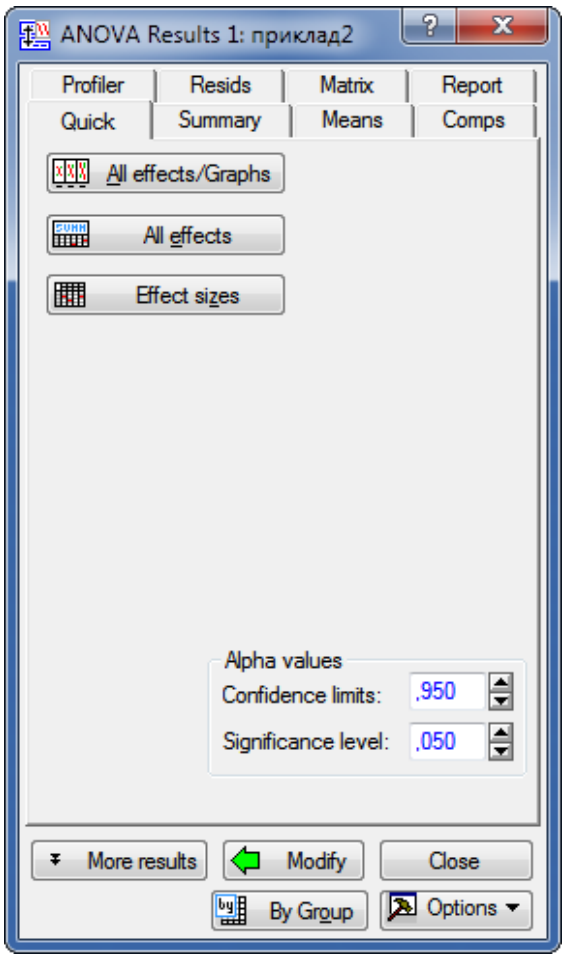


Рис. 6.3. Діалогове вікно подання результатів аналізу ***Factorial ANOVA***

Дана команда викликає діалог ***Table of All Effects*** (рис. 6.4), що містить результати і використовується для перегляду вибраних з даної таблиці ефектів у вигляді графіків середніх або таблиць. У даній таблиці будуть подані такі результати: *SS* – сума квадратів відхилень від середньої по групі (дисперсія), *D…* – ступені вільності, *MS* – середній квадрат (сума квадратів поділена на ступінь свободи), *F* – критерій Фішера, *p* – рівень значущості. Значущі ефекти ($p<0,05$) в таблиці ***Table of All Effects*** помічені зірочкою (*).

Виходячи з даних діалогового вікна (рис. 6.4) можна зробити такі висновки: $478 \cdot 10^2$ – частина загальної варіації, що формується під впливом фактору “Категорія пасажирів”, $606 \cdot 10^2$ – частина загальної варіації, що формується під впливом фактору “Вид транспорту” та $114 \cdot 10^3$ – частина загальної варіації, що формується під обох впливом обох факторів.

Якщо вибрати ***Spreadsheet (Таблиця)*** в рамці ***Display (Відображати)***, необхідну змінну (можна двічі клацнути) та натиснути ***OK***, то з’явиться таблиця (рис. 6.5) із значеннями середніх всіх залежних змінних і інших статистик в групах, відповідних чинникам (трьом рівням категоріального предиктора).

У рамці ***Means*** можна вказати різні способи обчислення середніх. ***Unweighted (Спостережувані незважені)*** середні обчислюються за допомогою усереднювання середніх за рівнями і комбінаціями рівнів чинників, які не використовувалися в таблиці (або графіках), після цього отримане значення ділиться на кількість середніх. ***Weighted (Спостережувані зважені)*** середні обчислюються як стандартні середні значення для відповідних комбінацій рівнів чинників. ***Least squares (Середні найменші квадрати)*** є очікуваними середніми у генеральній сукупності для поточної моделі.

Щоб провести графічний аналіз, необхідно повернутися у вікно ***Table of All Effects***, виділити опцію ***Graph (Графік)*** в рамці ***Display***, вибрати змінну та натиснути ***OK***. Програма побудує графік середніх за обраною змінною (рис. 6.6).

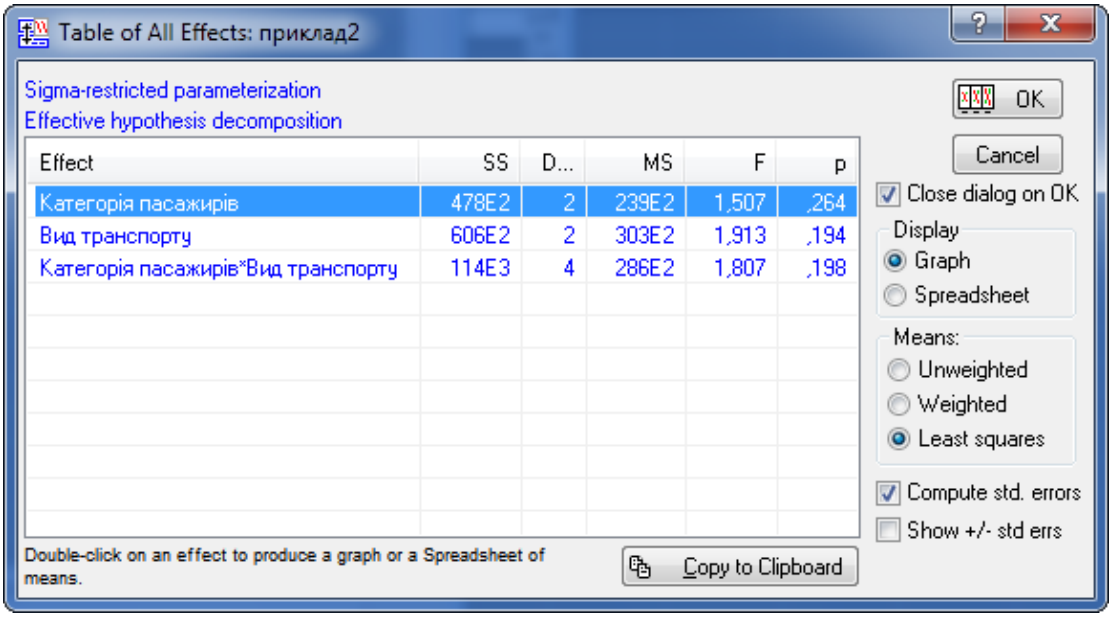


Рис. 6.4. Таблиця всіх ефектів

Вид транспорту; LS Means (приклад2)
Current effect: F(2, 11)=1,9126, p=,19371
Effective hypothesis decomposition

Cell No.	Вид транспорту	Витрати Mean	Витрати Std.Err.	Витрати -95,00%	Витрати +95,00%	N
1	Автобус	508,4667	54,70394	388,0641	628,8692	8
2	Маршрутне таксі	354,3889	56,80871	229,3538	479,4240	6
3	Метро	440,6667	56,80871	315,6315	565,7018	6

Рис. 6.5. Таблиця результатів

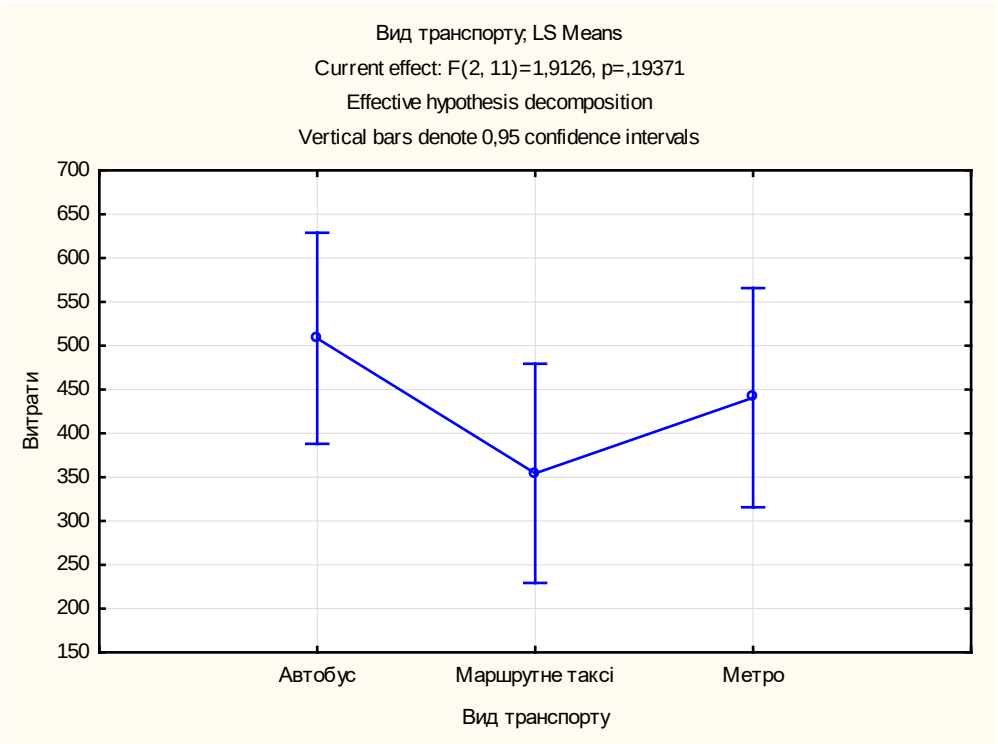


Рис. 6.6. Графічне подання результатів

Щоб проаналізувати вплив одразу двох факторів, необхідно обрати рядок з іменами обох змінних, вибрати бажане відображення результатів в рамці ***Display*** і натиснути ***OK***. Якщо залежних змінних дві або більше, слід вказати ім'я потрібної змінної, у вікні ***Dependent vars for the... (Залежні змінні для…)***. У вікні ***Arrangement of Factors (Розташування чинників)*** можна вказати порядок вибору взаємодіючих чинників в полях ***x-axis, upper (Вісь x, верх)*** і ***Line pattern (Шаблон лінії)*** (рис. 6.7.) Після натискання кнопки ***OK***, з'являться графіки середніх (рис. 6.8).

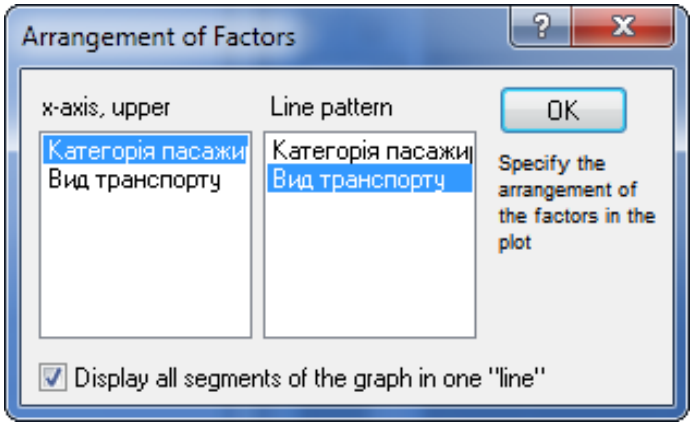


Рис. 6.7. Вибір розташування чинників

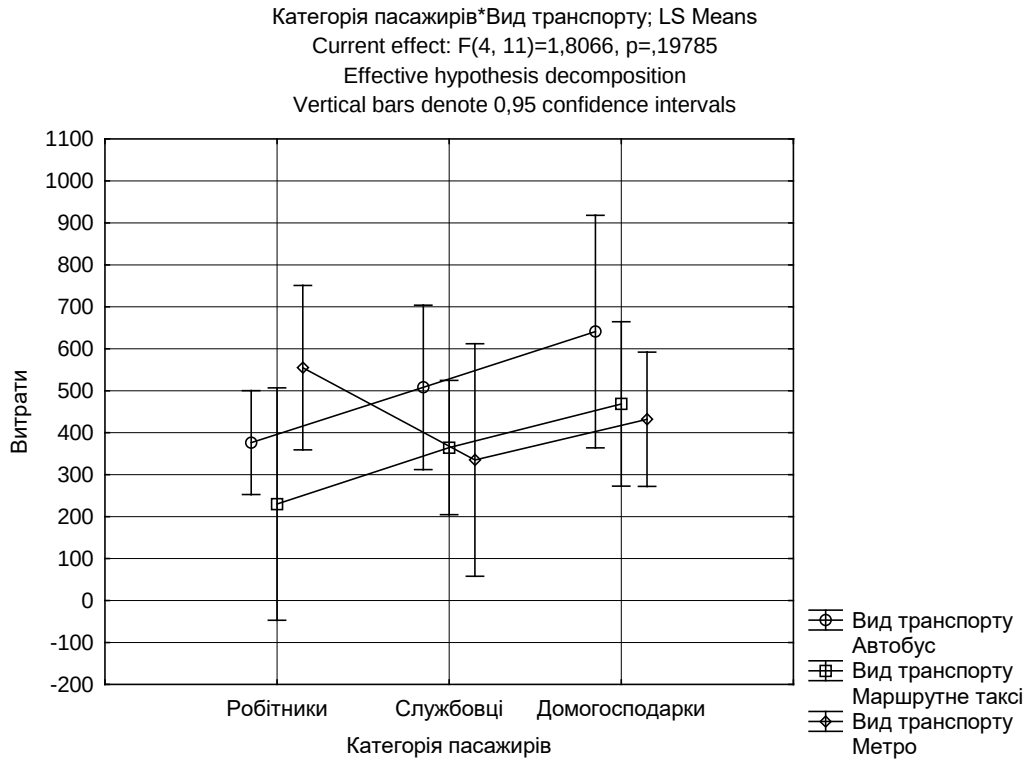


Рис. 6.8. Графіки середніх

Однак розглянутий варіант дисперсійного аналізу є параметричним, тобто передбачає виконання наступних обов'язкових умов щодо вхідних статистичних даних: 1) у кожній з порівнюваних груп значення аналізованої ознаки розподіляються нормально; 2) групові дисперсії однорідні (тобто між ними немає статистично значущої різниці). Крім того, всі порівнювані вибірки повинні бути незалежними. Тому перед отриманням результатів аналізу слід перевірити, чи виконуються зазначені умови, і чи коректно, використовуючи даний метод дисперсійного аналізу.

Для перевірки умов ***ANOVA*** необхідно виконати наступне:

1. Натиснути на кнопку ***More results (Додаткові результати)***, розташовану в нижній частині вікна ***ANOVA Results 1***. У результаті цього з'явиться вікно, подане на рис. 6.9.

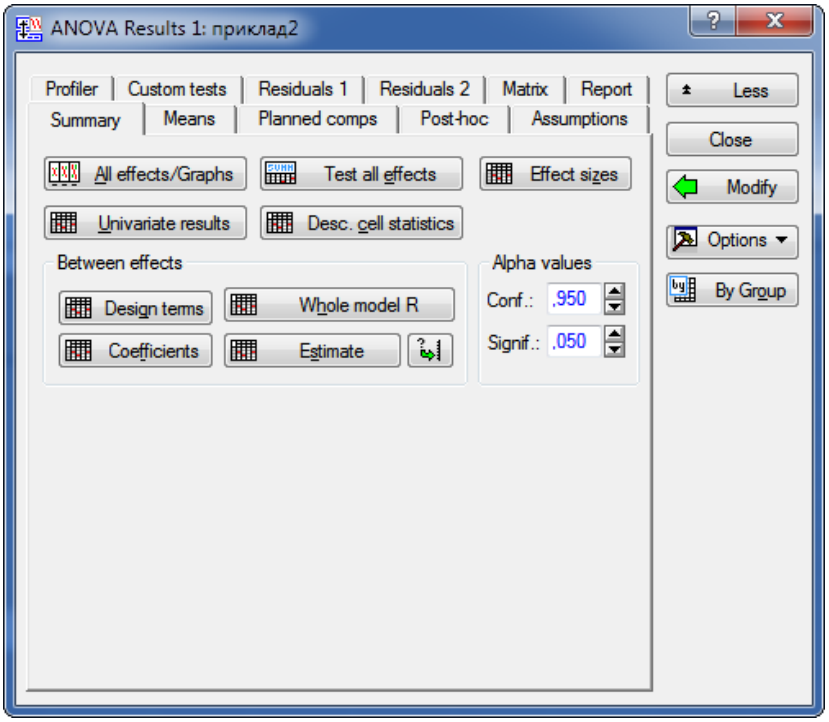


Рис. 6.9. Діалогове вікно вибору додаткового аналізу ***ANOVA***

2. Відкрити закладку ***Assumptions (Припущення)***. Для перевірки однорідності групових дисперсій в полі ***Homogeneity of variances/covariances (Однорідність дисперсій/коваріацій)*** натиснути на кнопку ***Levene's test (тест Левена)***. Якщо результат цього тесту вказує на відсутність відмінностей між дисперсіями ($p>0,05$), то застосування параметричного методу дисперсійного аналізу є обґрунтованим.

3. Для перевірки нормальності розподілу аналізованих даних необхідно скористатися однією з опцій, доступних в полі ***Distribution of variables within groups (Розподіл змінних усередині груп).*** Якщо число спостережень в порівнюваних групах невелика, краще використовувати графік нормальних ймовірностей. Якщо ж спостережень багато, то можна оцінити характер розподілу, побудувавши гістограми. При натисканні на одну з цих кнопок програма запропонує список груп, що беруть участь в аналізі.

4. Необхідно на вкладці ***Summary (Результати)*** натиснути кнопку ***Test all effects (Перевірити всі ефекти)***. У таблиці результатів, що з'явилася необхідно знайти комірки з величиною помилки p для нульової гіпотези про відсутність зв'язку між змінними. Якщо нерівність $p<<0,05$ виконується, то можна зробити висновок, що змінні статистично значимо розрізняється.

Для визначення значущості відмінності між середніми в групах потрібно використовувати апостеріорні порівняння для перевірки різниці середніх.

Для цього необхідно у діалозі ***ANOVA Results 1*** натиснути кнопку ***More results (Більше результатів)*** і у вікні, що відкрилося, вибрати вкладку ***Post-hoc (Апостеріорний аналіз)***, на якій подано різні апостеріорні критерії (рис. 6.10). Програма **STATISTICA** пропонує ряд тестів для множинних порівнянь: ***Fisher LSD***, ***Bonferroni***, ***Scheffe***, ***Tukey HSD***, ***Newman-Keuls***, ***Duncan's***, ***Dunnet***. Всі ці критерії дозволяють порівнювати середні за відсутності апріорної гіпотези щодо цих середніх.

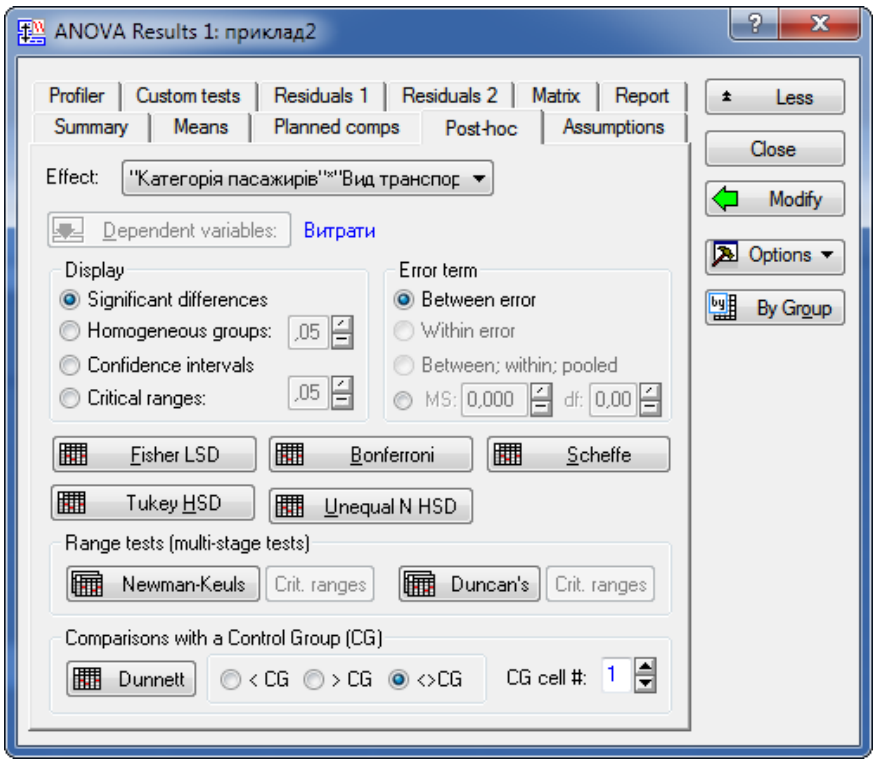


Рис. 6.10. Вкладка ***Post-hoc***

Проведемо тест найменшої значущої різниці (НЗР). Для цього необхідно вибрати залежну змінну, ефект і натисніть кнопку ***Fisher LSD (Тест Фішера найменшої значущої різниці)***. За допомогою НЗР оцінюється різниця між середніми. Якщо різниця d між будь-якими двома оцінками середнього перевищує або, принаймні, дорівнює НЗР, то середні значення відрізняються з імовірністю 1-α. У таблиці (рис. 6.11), що відкриється, в першому рядку наведені значення середніх, в стовпці 1 – назви груп, в решті комірок – рівні значущості. Нульова гіпотеза формулюється для двох середніх і стверджує, що ці середні рівні між собою. Червоним показані ті випадки, де нульова гіпотеза про рівність середніх відкидається.

LSD test; variable Витрати (приклад2)
Probabilities for Post Hoc Tests
Error: Between MS = 15843,, df = 11,000

Cell No.	Категорія пасажирів	Вид транспорту	{1} 376,40	{2} 230,00	{3} 555,00	{4} 508,00	{5} 364,67	{6} 335,00	{7} 641,00	{8} 468,50	{9} 432,00
1	Робітники	Автобус		0,311091	0,117975	0,237359	0,900733	0,769579	0,081285	0,400499	0,557539
2	Робітники	Маршрутне таксі	0,311091		0,058755	0,098760	0,374036	0,567195	0,041376	0,150102	0,192061
3	Робітники	Метро	0,117975	0,058755		0,715942	0,125839	0,181309	0,588103	0,506170	0,307328
4	Службовці	Автобус	0,237359	0,098760	0,715942		0,238138	0,285660	0,406686	0,759528	0,521945
5	Службовці	Маршрутне таксі	0,900733	0,374036	0,125839	0,238138		0,841988	0,083767	0,385531	0,525817
6	Службовці	Метро	0,769579	0,567195	0,181309	0,285660	0,841988		0,113580	0,404980	0,518272
7	Домогосподарки	Автобус	0,081285	0,041376	0,588103	0,406686	0,083767	0,113580		0,286983	0,178267
8	Домогосподарки	Маршрутне таксі	0,400499	0,150102	0,506170	0,759528	0,385531	0,404980	0,286983		0,756688
9	Домогосподарки	Метро	0,557539	0,192061	0,307328	0,521945	0,525817	0,518272	0,178267	0,756688	

Рис. 6.11. Таблиця результатів

3. Опис процедури Repeat measures ANOVA

Розглянемо дисперсійний аналіз з повторними вимірюваннями. На стартовій панелі ***General ANOVA/MANOVA*** (рис. 6.1) у списку ***Type of analysis*** необхідно вибрати ***Repeat measures ANOVA***; у списку ***Specification method – Quick Specs Dialog***. У результаті відкриється вікно діалогу (рис. 6.12) ***ANOVA/MANOVA Repeat measures ANOVA***.

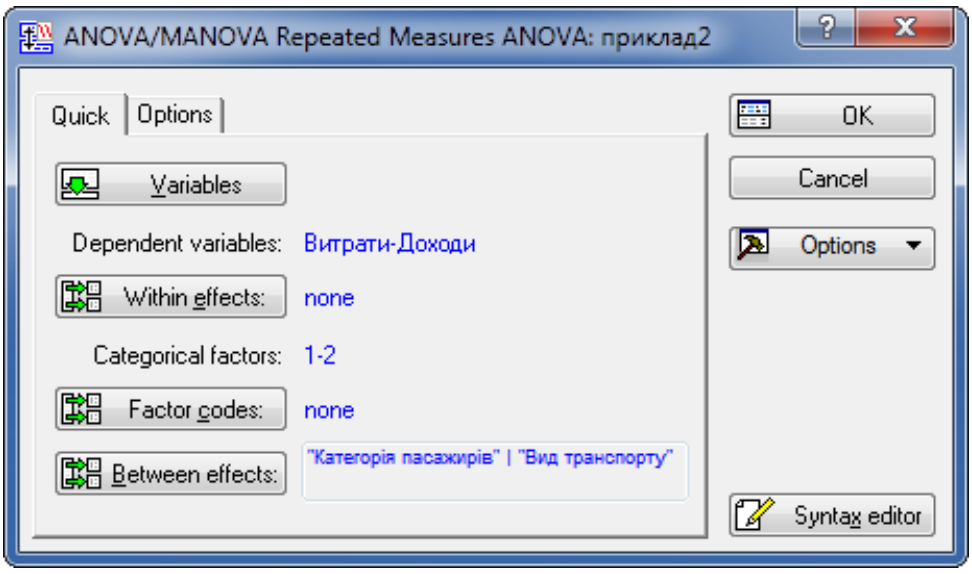


Рис. 6.12. Вибір змінних для ***ANOVA/MANOVA Repeat measures ANOVA***

Змінні необхідно вибрати на вкладці ***Quick***. Для введення фактору повторних вимірювань, потрібно вибрати кнопку ***Within effects (Внутрішньогрупові ефекти)***. Відкриється вікно (рис. 6.13) ***Specify within-subjects factor (Специфікація чинника повторних вимірювань)***. Дана процедура дозволяє ввести тільки один фактор (змінну, багато разів виміряну).

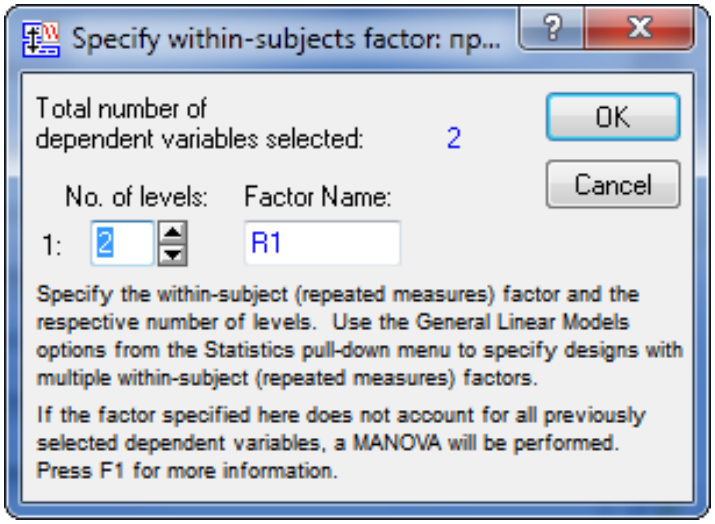


Рис. 6.13. Специфікація фактора повторних вимірювань

При необхідності проведення аналізу з великою кількістю факторів необхідно скористатися модулем ***GLM (Загальні лінійні моделі)***. ***No. of levels (Число рівнів)*** відповідає кількості повторних вимірювань. Можна змінити число рівнів і задати ім'я фактору в полі ***Factor Name.*** За допомогою кнопки ***Factor codes*** в діалозі ***ANOVA/MANOVA Repeat measures ANOVA*** можна задати коди рівнів категоріальних предикторів. У результаті натискання ***OK*** з'явиться вікно ***ANOVA Results 1***.

Подальший аналіз аналогічний описаному раніше.

4. Типовий приклад

Проаналізувати відмінність середньої врожайності культури залежно від типу добрива, виходячи з наступного: для перевірки ефективності двох нових добрив був виконаний такий експеримент. На дослідному полі випадковим чином були обрані 27 однакових за площею ділянок. Навесні в ґрунт 9 з них

внесли "старе" добриво, в ґрунт 8 – нове добриво 1, а в решту 10 – нове добриво 2. У кінці року була визначена урожайність культури, використаної в експерименті. Отримані дані наведені в таблиці 6.1.

Таблиця 6.1

№ з/п	Добриво	Врожайність	№ з/п	Добриво	Врожайність
1	Старе	1920	15	Нове 1	2090
2	Старе	2020	16	Нове 1	2252
3	Старе	2060	17	Нове 1	2360
4	Старе	1960	18	Нове 2	2320
5	Старе	1960	19	Нове 2	2240
6	Старе	2140	20	Нове 2	2100
7	Старе	1980	21	Нове 2	2296
8	Старе	1940	22	Нове 2	2249
9	Старе	1790	23	Нове 2	2321
10	Нове 1	2250	24	Нове 2	2368
11	Нове 1	2410	25	Нове 2	2205
12	Нове 1	2260	26	Нове 2	2261
13	Нове 1	2200	27	Нове 2	2400
14	Нове 1	2300			

Розв'язування. *Скористаємося* ***One-way ANOVA (Однофакторним дисперсійним аналізом)*** *(оскільки перевіряється вплив лише одного фактора - типу добрива). Запустимо модуль* ***One-way ANOVA*** *та виберемо змінні (рис. 6. 14).*

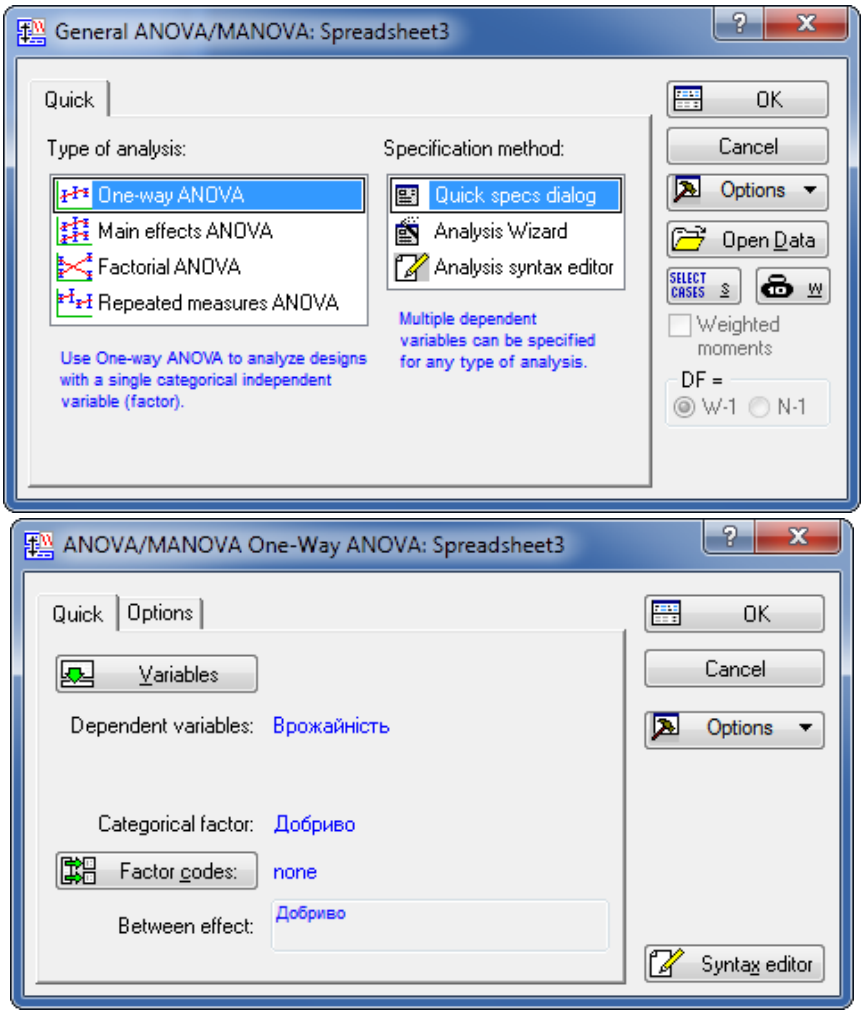


Рис. 6.14. Вибір методу аналізу та змінних

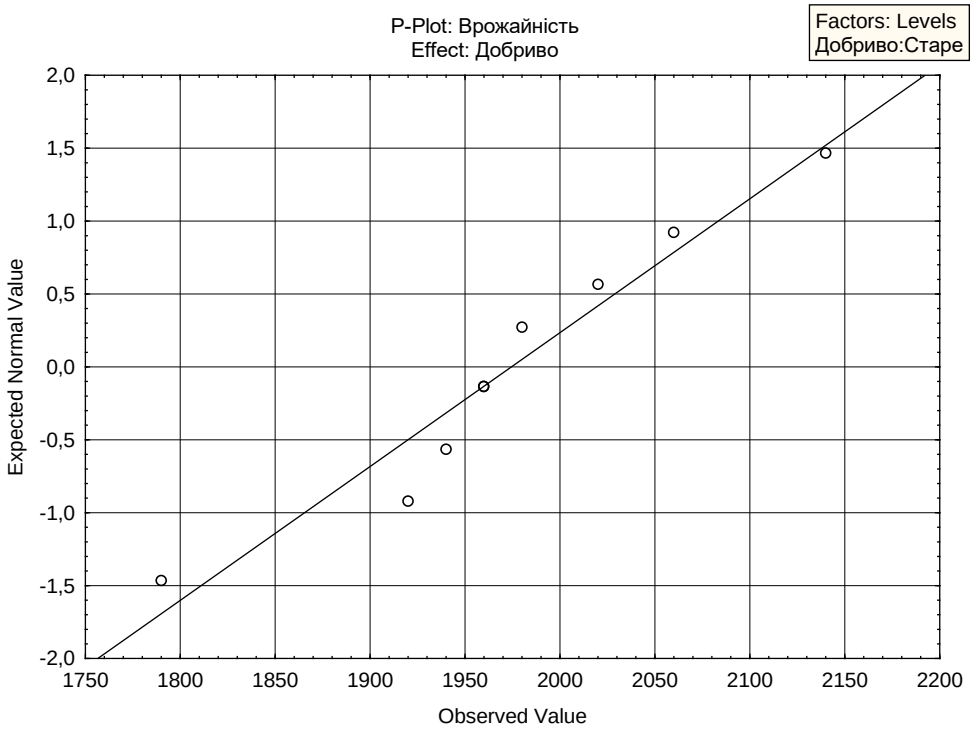
*Перевіримо умови **ANOVA**. Для цього у діалоговому вікні **ANOVA Results 1** виберемо **More results.** Перейдемо на вкладку **Assumptions** і перевіримо однорідність групових дисперсій (рис. 6.15).*

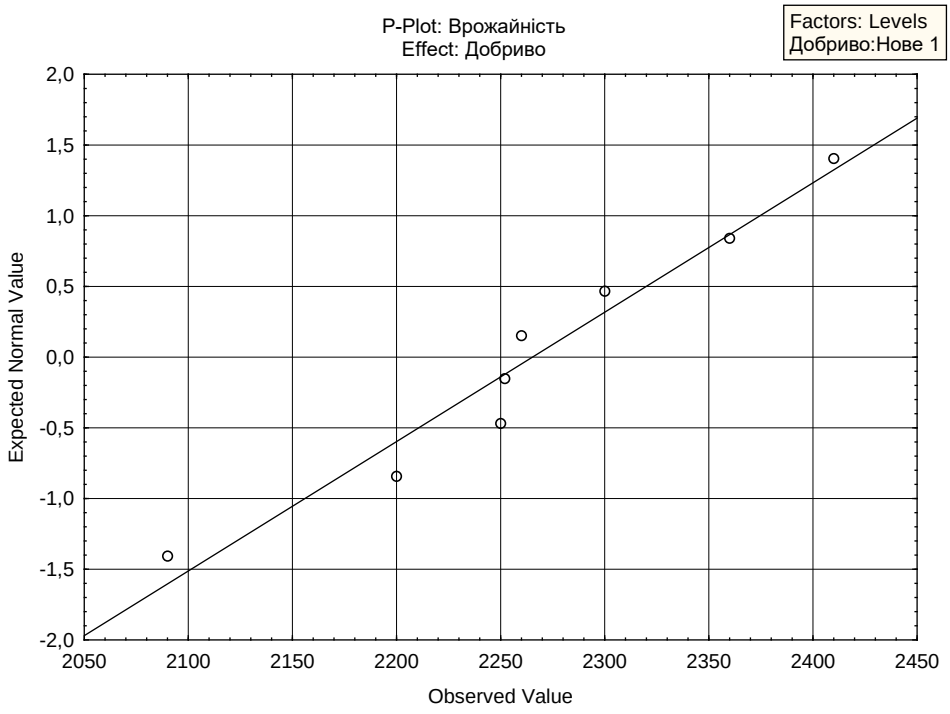
Levene's Test for Homogeneity of Variances (Spreadsheet3)
Effect: Добриво
Degrees of freedom for all F's: 2, 24

	MS Effect	MS Error	F	p
Врожайність	29,20247	3665,933	0,007966	0,992068

Рис. 6.15. Результат перевірки на однорідність

Як випливає з таблиці, різниці між дисперсіями немає (p>0,05).





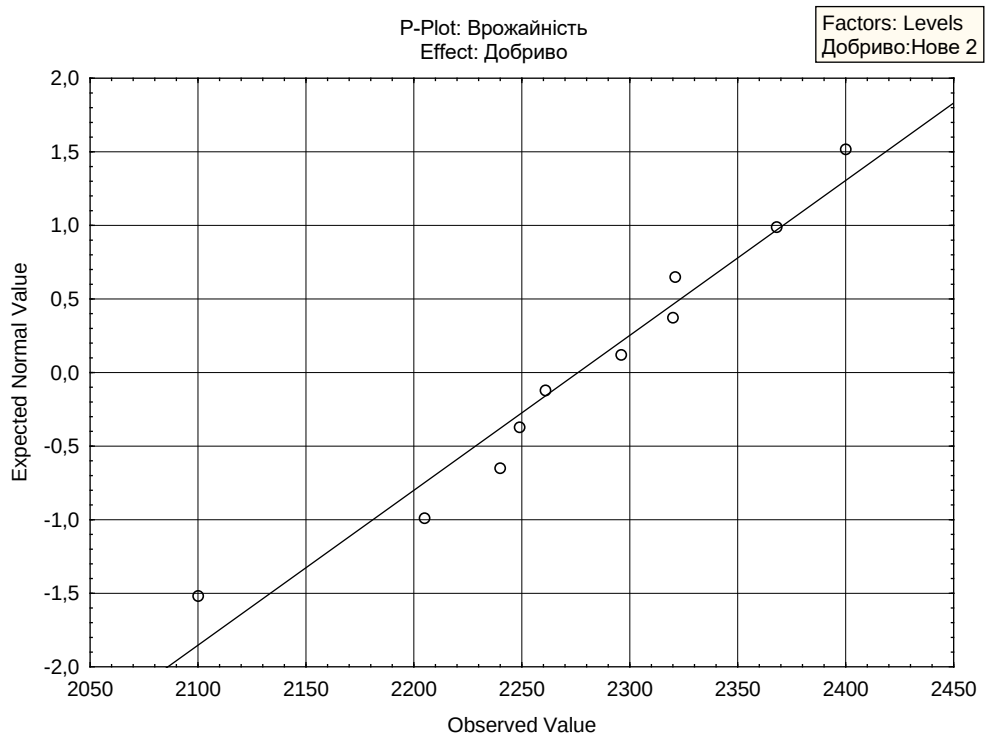


Рис. 6.16. Графіки нормальних розподілів

Перевіримо нормальність розподілу аналізованих даних. У полі ***Distribution of variables within groups*** *і побудуємо* ***Normal p-p (Графіки нормального розподілу)*** *(рис. 6.16). З графіків випливає, що точки спостереження тісно лежать уздовж теоретично очікуваної прямої. Отже, можна стверджувати, що розподіл змінних відповідає нормальному, а застосування дисперсійного аналізу до обраних змінних є обґрунтованим.*

Перейдемо до аналізу. На вкладці ***Quick*** *або* ***Summary*** *виконаємо команду* ***All effects/Graphs.*** *Таблиця всіх ефектів матиме вигляд, поданий на рис. 6.17. Очевидно (рис. 6.17), фактор значущий. Графічний аналіз (рис. 6.18) підтверджує, що середня врожайність значно підвищується з використанням нових добрив. Отже, зміна добрива суттєво впливає на середню врожайність (за використання старого добрива середня врожайність становить 1974,4 ц/га, нове 1 – 2265,25 ц/га та нове 2 – 2276 ц/га).*

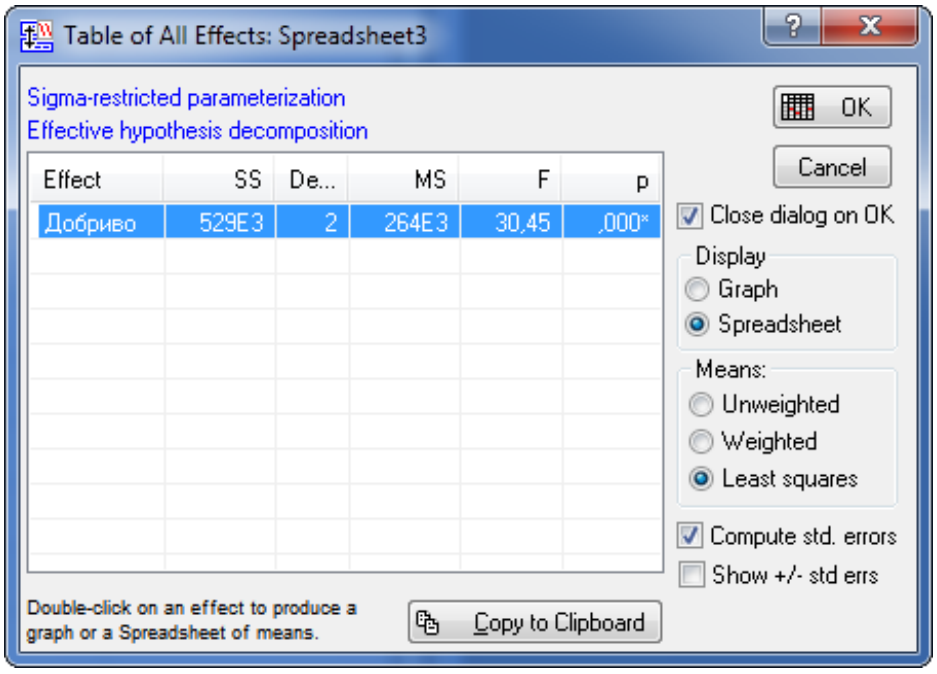


Рис. 6.17. Таблиця всіх ефектів

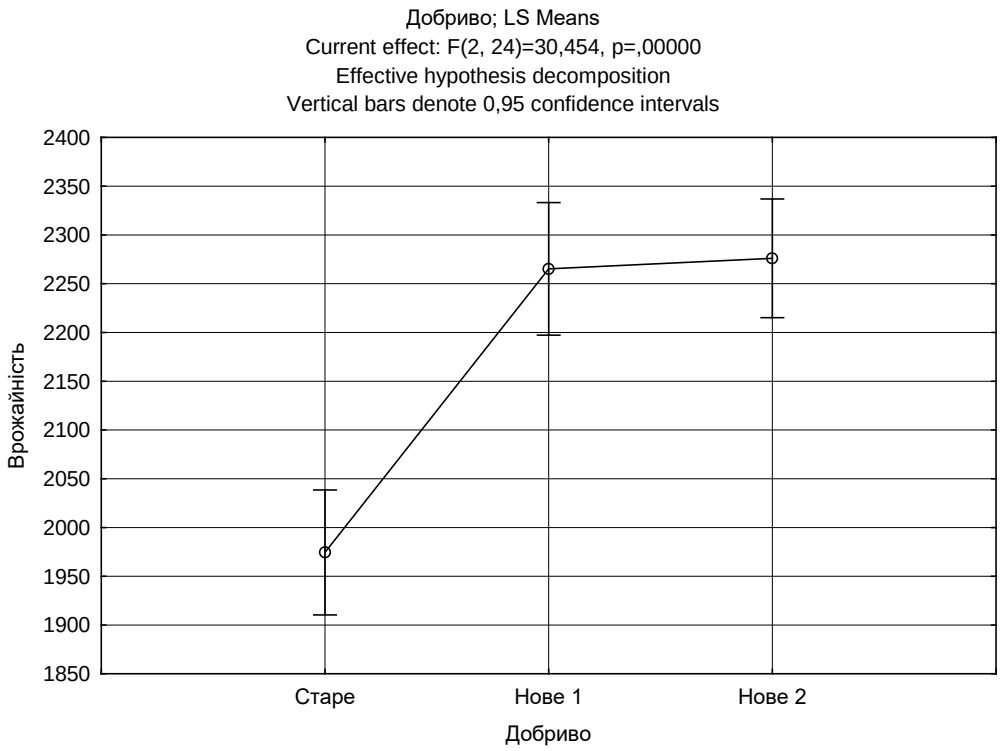


Рис. 6.18. Графічний аналіз впливу фактора

Завдання для самостійної роботи

6.1. Постачання продукції для деякої компанії здійснюються трьома постачальниками (“Мега +”, “Коста” і “Трамп”) в різний час: денні години, нічні зміни та навіть у перезмінки. Було проведено експертне оцінювання якості продукції, що постачається в різний час (табл. 6.2). Оцінити чи є відмінність у якості продукції, яка поставляється в різний час за даними поданими в таблиці? Якість продукції якого постачальника найвища?

Таблиця 6.2

Постачальники	*Оцінка якості, бали*		
	Денна зміна	*Нічна зміна*	*Перезмінка*
Мега+	77,06	93,12	77,05
Коста	81,14	88,13	78,11
Трамп	82,02	81,18	79,91

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх спостережень.

Data: 6_1 (3v by 9c)

	1 *Постачальники*	2 Зміна	3 *Оцінка якості, бали*
1	Мега+	Денна	77,06
2	Мега+	Нічна	93,12
3	Мега+	Перезмінка	77,05
4	Коста	Денна	81,14
5	Коста	Нічна	88,13
6	Коста	Перезмінка	78,11
7	Трамп	Денна	82,02
8	Трамп	Нічна	81,18
9	Трамп	Перезмінка	79,91

6.2. Потрібно оцінити вплив рівня реклами на обсяги продажів магазинів за даними таблиці 6.3.

Таблиця 6.3

Торговельна точка	*Рівень реклами*		
	високий	*середній*	*низький*
	Продажі, тис. грн.		
1	10	8	5
2	9	8	7
3	10	7	6
4	8	9	4
5	9	6	5
6	8	4	2
7	9	5	3
8	7	5	2
9	7	6	1
10	6	4	2

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх спостережень.

Data: 6_2 (3v by 30c)

	1 *Торговельна точка*	2 *Рівень реклами*	3 *Продажі, тис. грн.*
1	1	високий	10
2	1	середній	8
3	1	низький	5
4	2	високий	9
5	2	середній	8
6	2	низький	7
7	3	високий	10
8	3	середній	7
9	3	низький	6
10	4	високий	8

6.3. Оцінити вплив місця розташування магазину на обсяг реалізації за даними таблиці 6.4.

Таблиця 6.4

	Обсяг реалізації, тис. грн.						
	Пн.	Вт.	Ср.	Чт.	Пт.	Сб.	Нд.
Центр	14	15	14,8	15	15,2	15	15,4
Південно-кільцева	14,4	14,9	14,9	15,5	14,3	14,5	14,7
Проспект Незалежності	14,2	15,2	14,6	15,4	15,2	15,0	14,8

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх спостережень.

Data: 6_3 (3v by 21c)

	1 **Місце розташування**	2 День	3 *Обсяг реалізації, тис. грн.*
1	Центр	ПН	14
2	Центр	ВТ	15
3	Центр	СР	14,8
4	Центр	ЧТ	15
5	Центр	ПТ	15,2
6	Центр	СБ	15
7	Центр	НД	15,4
8	Південно-кільцева	ПН	14,4
9	Південно-кільцева	ВТ	14,9
10	Південно-кільцева	СР	14,9

6.4. У таблиці 6.5 подано дані про виробіток за день залежно від розподілу робітників за загальним стажем. Оцінити вплив стажу на кількість виробленої продукції. Проаналізувати виробіток за цехами.

Таблиця 6.5

Групи робітників за стажем (років)	*Вироблено продукції, шт. в день*		
	Цех № 1	*Цех № 2*	*Цех № 3*
0-5	30	70	120
5-10	50	50	180
10-15	80	40	110
15-20	100	20	30
20-25	30	10	40
25-30	10	20	20

Примітка. Файл для аналізу у STATISTICA формується наступним чином.

Data: 6_4 (3v by 18c)

	1 ***Групи робітників за стажем (років)***	2 *Цех*	3 *Вироблено продукції, шт. в день*
1	0-5	1	30
2	0-5	2	70
3	0-5	3	120
4	5-10	1	50
5	5-10	2	50
6	5-10	3	180
7	10-15	1	80
8	10-15	2	40
9	10-15	3	110
10	15-20	1	100

6.5. У таблиці 6.6 подано дані про роботу фірми-туроператора за 2011 рік.

Таблиця 6.6

Напрямок	*Місяць*	*Кількість туристів, чол.*
Країни Західної Європи	Грудень, Січень, Лютий, Червень, Липень, Серпень	98, 95, 102, 104, 95, 107
Країни Східної Європи	Травень, Червень, Липень, Серпень	85, 88, 98, 100
Єгипет	Січень, Лютий, Березень	104, 101, 103
Туреччина	Березень, Квітень, Травень, Червень, Липень, Серпень	112, 115, 117, 120, 125, 110

Оцінити вплив напрямів подорожі на кількість туристів. Оцінити вплив місяця на кількість туристів та вплив обох факторів одночасно. Провести дисперсійний аналіз, вважаючи, що змінна місяць виміряна двічі.

Примітка. Файл для аналізу у STATISTICA формується наступним чином.

Data: 6_5 (10v by 22c)

	1 *Напрямок*	2 *Місяць*	3 *Кількість туристів, чол.*
1	Країни Західної Європи	Грудень	98
2	Країни Західної Європи	Січень	95
3	Країни Західної Європи	Лютий	102
4	Країни Західної Європи	Червень	104
5	Країни Західної Європи	Липень	95
6	Країни Західної Європи	Серпень	107
7	Країни Східної Європи	Травень	85
8	Країни Східної Європи	Червень	88
9	Країни Східної Європи	Липень	98
10	Країни Східної Європи	Серпень	100

Лабораторна робота № 7
Закони розподілу

1. Основні теоретичні відомості законів розподілу

У теорії статистики при розв'язуванні певних задач, пов'язаних з перевіркою статистичних гіпотез, дисперсійним, кореляційним, регресійним та іншими видами статистичного аналізу, необхідно спочатку визначати закони розподілу.

Аналіз статистичних даних у вигляді варіаційних рядів припускає виявлення закономірностей розподілу, визначення та побудову деякої теоретичної форми розподілу. Крива розподілу може розглядатися як деяка теоретична (ймовірнісна) форма розподілу, яка властива деякій сукупності в конкретних умовах. Отже, аналізуючи частоти в емпіричному розподілі, можна описати розподіл за допомогою математичної моделі – закону розподілу, встановити за вихідними даними параметри теоретичної кривої і перевірити правильність висунутої гіпотези про тип розподілу даного ряду даних.

При дослідженні закономірностей розподілу дуже важливо висунути деяку гіпотезу про тип кривої розподілу, оскільки, якщо крива описана математично правильно, то вона більш точно відображає закономірності цього розподілу і може бути використана в практичних розрахунках.

Крива розподілу називається емпіричною, якщо вона є графічним зображенням у вигляді неперервної лінії зміни частот, функціонально пов'язаних зі зміною варіант. Емпіричні криві розподілу, побудовані на основі, як правило, невеликого числа спостережень, дуже важко описати аналітично, тому для виявлення статистичних закономірностей порівняння й узагальнення різних сукупностей аналогічних даних використовуються теоретичні розподіли.

Під теоретичним розподілом розуміється гіпотетичний розподіл імовірностей, який властивий для спостережуваних частот варіаційного ряду. Теоретичні розподіли – це добре вивчені розподіли, що є залежностями між щільністю розподілу і значеннями ознаки та відображають закономірності розподілу. Вони описуються статистичними функціями, параметри яких обчислюються за певними характеристиками сукупності, що вивчається.

Дослідження форми розподілу припускає заміну емпіричного розподілу відомим теоретичним розподілом, близьким йому за формою. При цьому необхідно дотримуватися умови: відмінності між емпіричним і теоретичним розподілами повинні бути мінімальними. Це означає, що сума частот емпіричного розподілу повинна відповідати сумі частот теоретичного розподілу, тобто $\sum n_{емп} \approx \sum n_{теор}$.

Теоретичний розподіл у цьому випадку є деякою моделлю емпіричного розподілу, що ідеалізується, й аналіз варіаційного ряду зводиться до зіставлення емпіричного і теоретичного розподілів і визначення відмінностей між ними.

У практиці статистичних досліджень використовуються наступні теоретичні розподіли (закони розподілу). Нормальний розподіл – найважливіший закон розподілу неперервних випадкових величин. За допомогою нормального розподілу можна описати більшість явищ навколишнього світу, він використовується для моделювання багатьох економічних процесів. Розподіл χ^2 (Пірсона) асиметричний, має позитивну правобічну асиметрію та відіграє важливу роль при перевірці залежностей у таблицях спряженості і критеріях згоди. Розподіл Стьюдента (*t*-розподіл) застосовується при оцінці середнього, в регресійному аналізі, при використанні часових рядів. *F*-розподіл Фішера асиметричний, має позитивну правобічну асиметрію та використовується при оцінці дисперсії випадкової величини, в

регресійному, дисперсійному і дискримінантному аналізі, а також в інших видах багатовимірного аналізу даних. Логарифмічно-нормальний розподіл використовується для опису розподілу доходів, банківських вкладів, місячної заробітної плати, посівних площ під різні культури, довговічності виробів в режимі зносу і старіння тощо. Біноміальний розподіл широко використовується в теорії і практиці статистичного контролю якості продукції, при моделюванні систем масового обслуговування, в теорії стрільби та інших областях. Розподіл Пуассона є наближенням біноміального розподілу при достатньо великих і малих значеннях імовірності та використовується при моделюванні числа збоїв на автоматичній лінії, відмов складної системи тощо. Розподіл Бернуллі описує ситуації, де результатами є успіх або невдача. Геометричний розподіл використовують тоді, коли моделюють ситуації, в яких випробування проводяться до першого настання успіху.

У системі STATISTICA існує модуль ***Probability calculator (Ймовірнісний калькулятор)***, за допомогою якого можна знайти різні характеристики законів розподілу. Використовуючи його, можна також будувати графіки функції щільності та функції розподілу, для неперервних випадкових величин – обчислити відсоткові точки, визначити ймовірність попадання значень у заданий інтервал, для дискретних випадкових величин – обчислити ймовірності та будувати ряди розподілу.

Розглянемо основні принципи роботи даного модуля. Для його запуску необхідно у вкладці ***Statistics*** у групі ***Base*** в модулі ***Basic Statistics/Tables*** вибрати ***Probability calculator*** або в меню ***Statistics*** у модулі ***Basic Statistics/Tables*** вибрати команду ***Probability calculator***. Відкриється діалогове вікно ***Probability Distribution Calculator (Калькулятор імовірнісних розподілів)*** (рис. 7.1).

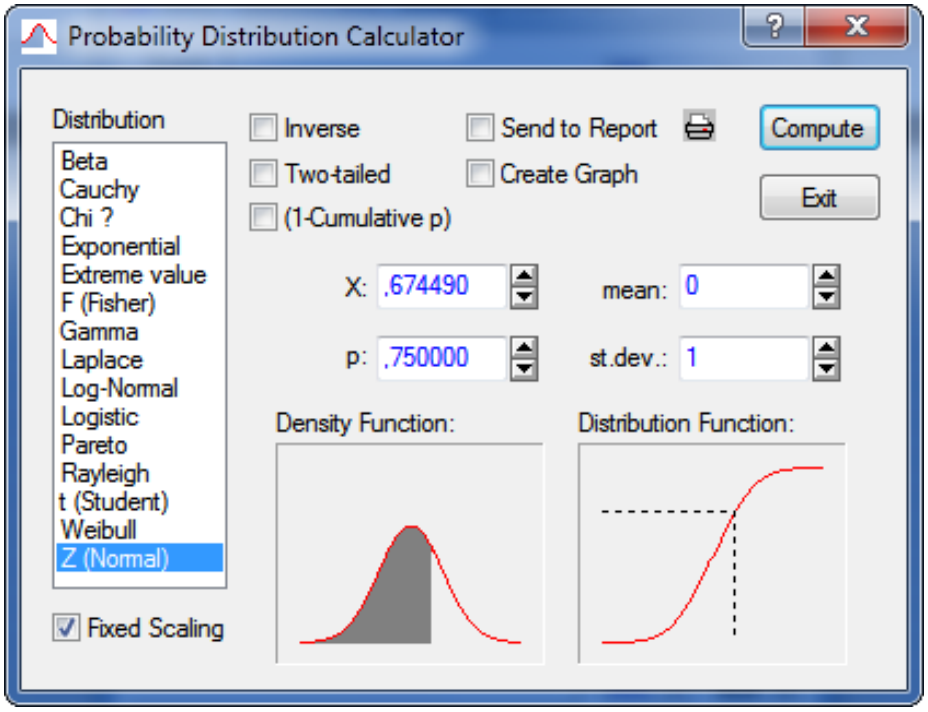


Рис. 7.1. Діалогове вікно ***Probability Distribution Calculator***

У лівій частині діалогового вікна розташовано список розподілів ***Distribution (Розподіл)***. При виборі закону розподілу праворуч автоматично з'являться поля, де можна задати характеристики розподілу, наприклад mean (середнє) і std.dev (середньоквадратичне відхилення) для нормального розподілу. За замовчуванням система запише в них стандартні значення, наприклад середнє = 0, середньоквадратичне відхилення = 1. Дані значення можна

змінити. Одночасно з вибором розподілу в лівому списку праворуч у діалоговому вікні з'являються графіки щільності та функції розподілу: ***Density Function (Функція щільності)***, ***Distribution Function (Функція розподілу)***. У полі *p* треба задати рівень ймовірності, при цьому позначка автоматично встановлюється на ***Inverse (Інверсія)***. Після натискання на кнопку ***Compute (Підрахунок)*** (у правому верхньому куті діалогового вікна) в полі X (для нормального розподілу) з'явиться значення квантиля, відповідне обраному рівню ймовірності. Можна і за заданим значенням X обчислити рівень імовірності *p*. Для цього треба задати значення квантиля, клацнути по кнопці ***Compute***; у полі *p* з'явиться значення ймовірності, відповідного значенню X. Якщо встановити позначку на ***Create Graph (Створити графік)*** і натиснути кнопку ***Compute***, то на екрані з'являться графіки щільності та функції розподілу з виділеними на них значеннями ймовірності і квантилів.

Якщо встановити позначку на ***Two-tailed (Подвійний критерій)*** на діалоговому вікні рис. 7.1, то розрахунок буде проведений для відрізка $[m-x; m+x]$ (де m – середнє значення), в іншому випадку – для відрізка $[-\infty; x]$. Якщо встановити позначку на ***1-Cumulative p (1-сукупний p)***, то розрахунок буде проведений для відрізка, протилежному вказаного. Позначка на ***Fixed Scaling (Фіксована шкала)*** під списком розподілів ***Distributions*** указує, що обрана фіксована шкала.

2. Підбір закону розподілу

При аналізі статистичних даних іноді виникає необхідність апроксимувати емпіричний розподіл відомим теоретичним законом розподілу. Для цієї мети в **STATISTICA** призначений модуль ***Distribution Fitting (Підгонка розподілу)***. Щоб запустити модуль ***Distribution Fitting***, необхідно у вкладці ***Statistics*** у групі ***Base*** або в головному меню ***Statistics*** вибрати однойменну команду. У вікні ***Distribution Fiting*** (рис. 7.2) треба вказати тип випадкової величини, тобто ***Continuous Distribution (Неперервна)*** або ***Discrete Distribution (Дискретна)***, а також гіпотетичний закон розподілу.

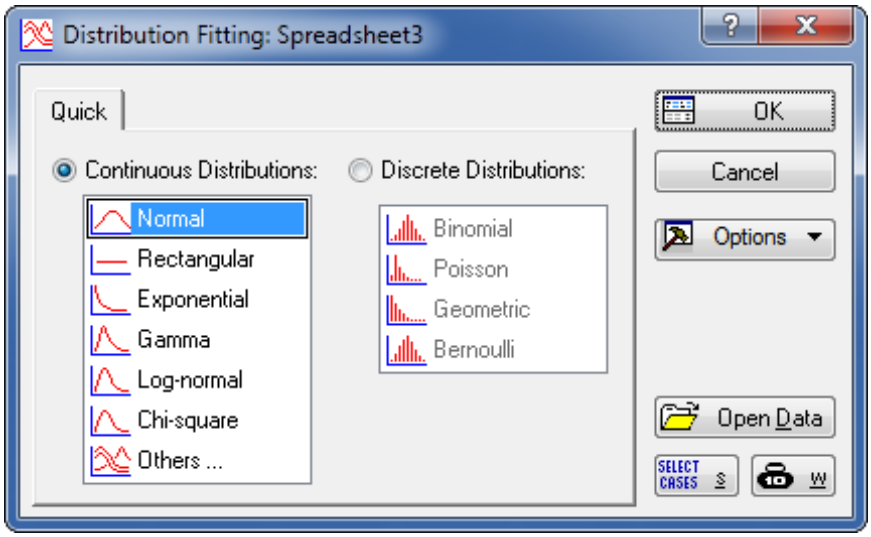


Рис. 7.2. Діалогове вікно ***Distribution Fiting***

Для неперервних випадкових величин запропоновано шість законів розподілу, а для дискретних – чотири. Необхідно вказати потрібний закон розподілу і натиснути ***OK***. З'явиться ще одне діалогове вікно ***Fitting Continuous Distributions (Підбір неперервного розподілу)*** або ***Fitting Discrete Distributions (Підбір дискретного розподілу)*** (залежно від обраного теоретичного закону розподілу у попередньому діалозі) (рис. 7.3). У діалоговому

вікні ***Fitting Continuous Distributions*** (рис. 7.3) вибирається змінна (кнопка ***Variables***). У лівому верхньому куті діалогового вікна стає активним спадний список і можна вибрати інший закон розподілу. За замовчуванням активується вкладка ***Quick***. На цій вкладці є дві кнопки: ***Summary: Observed and expected distributions (Результат: спостережувані і очікувані розподіли)*** і ***Plot of Observed and expected distributions (Графік спостережуваних та очікуваних розподілів)***.

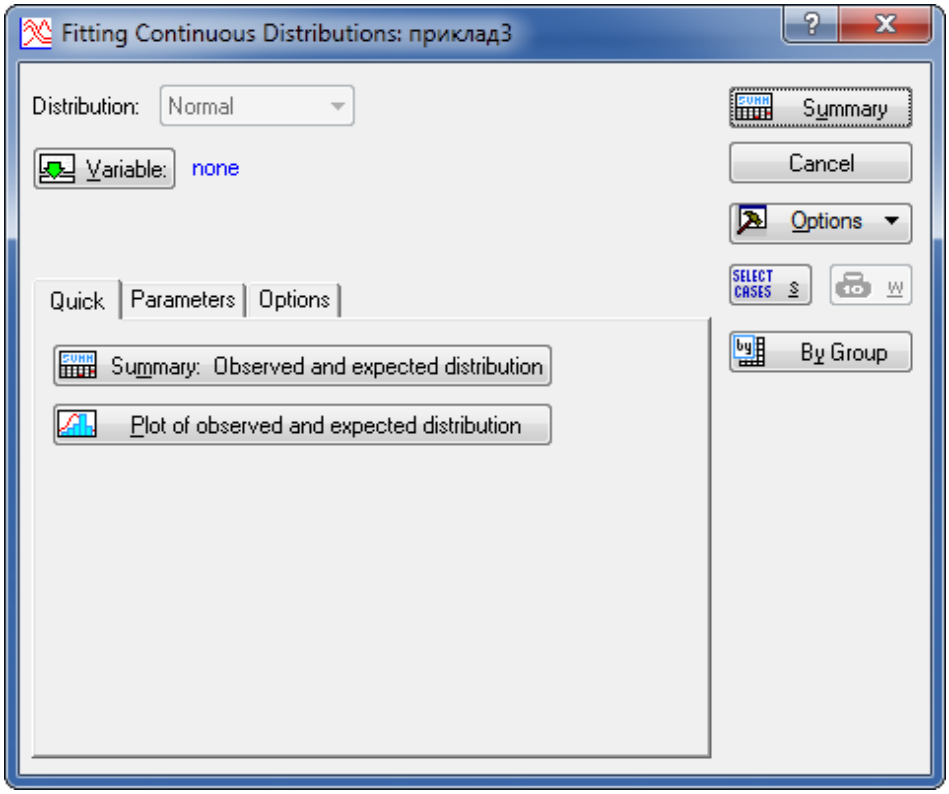


Рис. 7.3. Діалогове вікно ***Fitting Continuous Distributions***

Якщо натиснути на першу кнопку, то програма відобразить числові характеристики у вигляді таблиці (рис. 7.4). Кожний рядок цієї таблиці характеризує інтервал, у який потрапляють значення досліджуваної змінної. У першому стовпці ***Observed Frequency (Спостережувана частота)*** для кожного розгляненого інтервалу вказано кількість значень, що потрапили в цей інтервал. У другому стовпці ***Cumulative Observed (Сукупна спостережувана частота)*** для кожного інтервалу наведено кількість значень, що потрапили в цей і всі попередні інтервали (накопичені частоти). У третьому і четвертому стовпцях ***Percent Observed (Спостережуваний відсоток***) і ***Cumul. % Observed (Сумарний спостережуваний відсоток)*** указано ті ж величини, що і в попередніх двох, але обчислені у відсотках. У п'ятому стовпці ***Expected Frequency (Очікувані частоти)*** вказані частоти, що відповідають вибраному розподілу. В останньому стовпці ***Observed-Expected (Спостережувана частота – Очікувана частота)*** подана різниця між частотами емпіричного та теоретичного розподілів.

При натисканні на другу кнопку ***Plot of Observed and expected distributions*** діалогового вікна (рис. 7.3) буде побудована крива теоретичного закону розподілу та гістограма емпіричного розподілу (рис. 7.5). У заголовку гістограми вказано назва змінної, гіпотетичний закон розподілу, а також три числових параметри:

перший параметр – це значення критерію χ^2 (чим менше це значення, тим більша ймовірність того, що випадкова величина (досліджувана змінна) має вибраний гіпотетичний

закон розподілу);

другий параметр – *df* – число ступенів вільності;

третій параметр – *p* – рівень значущості критерію, який визначає ймовірність помилки при відхиленні гіпотези про нормальність.

Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
Variable: Врожайність, Distribution: Normal (Spreadsheet3) Chi-Square = 1,46131, df = 1 (adjusted) , p = 0,22672									
<= 1750,00000	0	0	0,00000	0,0000	0,164122	0,16412	0,60786	0,6079	-0,16412
1800,00000	1	1	3,70370	3,7037	0,201185	0,36531	0,74513	1,3530	0,79882
1850,00000	0	1	0,00000	3,7037	0,385979	0,75129	1,42955	2,7825	-0,38598
1900,00000	0	1	0,00000	3,7037	0,678462	1,42975	2,51282	5,2954	-0,67846
1950,00000	2	3	7,40741	11,1111	1,092654	2,52240	4,04687	9,3422	0,90735
2000,00000	3	6	11,11111	22,2222	1,612264	4,13467	5,97135	15,3136	1,38774
2050,00000	1	7	3,70370	25,9259	2,179650	6,31432	8,07278	23,3864	-1,17965
2100,00000	3	10	11,11111	37,0370	2,699822	9,01414	9,99934	33,3857	0,30018
2150,00000	1	11	3,70370	40,7407	3,063953	12,07809	11,34798	44,7337	-2,06395
2200,00000	1	12	3,70370	44,4444	3,185868	15,26396	11,79951	56,5332	-2,18587
2250,00000	4	16	14,81481	59,2593	3,035096	18,29906	11,24109	67,7743	0,96490
2300,00000	5	21	18,51852	77,7778	2,649206	20,94826	9,81187	77,5862	2,35079
2350,00000	2	23	7,40741	85,1852	2,118641	23,06690	7,84682	85,4330	-0,11864
2400,00000	3	26	11,11111	96,2963	1,552376	24,61928	5,74954	91,1825	1,44762
2450,00000	1	27	3,70370	100,0000	1,042158	25,66144	3,85984	95,0424	-0,04216
< Infinity	0	27	0,00000	100,0000	1,338564	27,00000	4,95764	100,0000	-1,33856

Рис. 7.4. Таблиця результатів підбору розподілу

У системі **STATISTICA** буквою *p* позначається статистична значущість (тобто рівень значущості для перевірки нульової гіпотези). Як правило, якщо $p \geq 0.05$, H_0 приймається, якщо $p < 0.05$, H_0 відкидається. Однак значення *0,05* можна змінювати, виходячи з цілей дослідження.

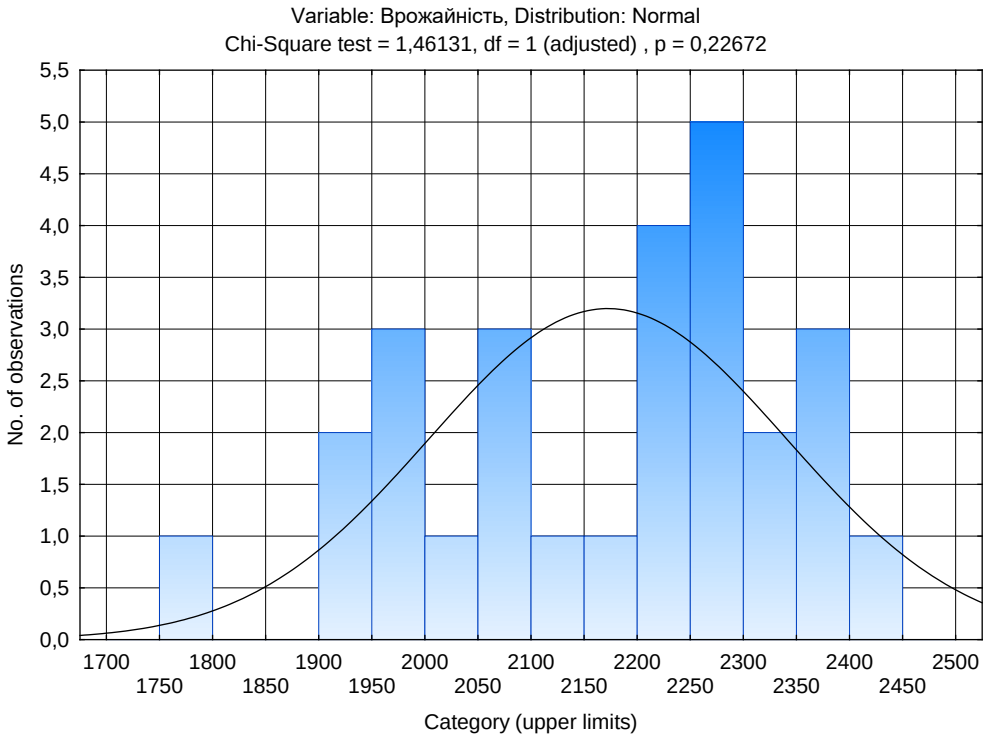


Рис. 7.5. Графік емпіричного та теоретичного розподілів

На вкладці ***Parameters (Параметри)*** діалогового вікна ***Fitting Continuous Distributions*** наведені значення параметрів гіпотетичного закону розподілу:

- ***Number of categories (Кількість категорій)*** – кількість інтервалів, на яку буде розбита вибіркова сукупність;

- ***Lower Limit, Uper Limit (Нижня і верхня межі)*** – за замовчуванням береться мінімальне та максимальне значення вибіркової сукупності відповідно (проте змінивши ці параметри, можна виключити з розгляду всі значення, які не потрапляють у досліджуваний інтервал);
- ***Mean and Variance (Середнє та дисперсія)*** – тільки для нормального розподілу. Ці параметри визначаються програмою автоматично, але їх можна зазначити і вручну, якщо, наприклад, потрібно не визначити закон розподілу, а перевірити, наскільки розподіл випадкової величини відрізняється від закону розподілу із заданими параметрами. Якщо обрані параметри не влаштовують, можна натиснути кнопку ***Set to default (Встановити значення за замовчуванням)***.

На вкладці ***Options (Опції)*** діалогового вікна ***Fitting Continuous Distributions*** відображено:

- ***Kolmogorov-Smirnov test (Критерій Колмогорова-Смірнова)*** для перевірки гіпотези про відповідність вибіркових даних тому чи іншому закону розподілу. Статистика Колмогорова-Смірнова дорівнює максимальній абсолютній різниці між гіпотетичною й емпіричної функціями розподілу. У цьому полі можна вибрати:
  - ✓ ***No*** (тест не обчислюється);
  - ✓ ***Yes (categorized)*** обчислюється за згрупованими (інтервальними) даними;
  - ✓ ***Yes (continuous)*** обчислюється за незгрупованими даними.
- ***Chi-Square test (Критерій χ^2 (Пірсона))*** для перевірки гіпотези про відповідність вибіркових даних тому чи іншому закону розподілу. Якщо в інтервал потрапило менше 5 значень, при установці прапорця на ***Combine Categories (Комбінувати категорії)*** він об'єднується з сусіднім до тих пір, поки кількість значень в інтервалах буде не менше 5.
- ***Graph***
  - ✓ ***Plot Distributions (Графіки розподілу)*** – вибирається тип графіку: ***Frequency distribution*** – графік щільності розподілу; ***Cumulative distributions*** – графік функції розподілу.
  - ✓ ***Plot row frequencies or % (Ряди частот чи відсотків)***: ***Raw frequencies*** – на вертикальній осі графіка будуть відкладені значення відносних частот; ***Plot raw frequencies or %*** – відсоткові значення.

3. Генерація випадкових чисел

У системі **STATISTICA** є можливість генерувати випадкові числа, що підпорядковуються певним законам розподілу. Відомо, що зі збільшенням обсягу вибірки зростає відповідність емпіричного розподілу теоретичному. Наприклад, якщо кількість чисел, що генеруються більше 1000, то відхилення емпіричного закону від теоретичного практично непомітне. Для генерації випадкових чисел треба двічі клацнути в таблиці даних (у яку передбачається записати згенеровані числа) на імені змінної. У вікні специфікації змінної натисніть кнопку ***Functions (Функції)***. У відкритому вікні треба виділити ***All Functions (Всі функції)*** і вибрати потрібну функцію (рис. 7.6).

Розглянемо деякі функції. Генератор випадкових чисел, розподілених рівномірно на відрізку [0; 1], запускається формулою ***rnd (1)***. Оператор ***rnd (b–a)+a*** генерує числа, розподілені рівномірно на відрізку [a; b]. Вибірка, розподілена за заданим законом, генерується у файл заданням у полі ***Long name (label, or formula with Functions)*** відповідного виразу:

= ***rnd (5)*** для R [0; 5];

= ***VNormal (rnd (1); 2; 3)*** для N (2; 3);

= ***VExpon (rnd (1); 1/2)*** для E (0, 5) з середнім $\mu = 1/2$;

= ***VCauchy (rnd (1); 0; 1)*** для C (0; 1);

= ***VLognorm (rnd (1); 0,5; 0,5)*** для Lgn (0,5; 0,5);

= ***VChi2 (rnd (1); 8)*** для χ^2,

де ***R*** – рівномірний, ***N*** – нормальний, ***E*** – експоненційний, ***C*** – Коші, ***Lgn*** – логнормальний, χ^2 – хі-квадрат розподіли, буква ***V*** вказує на те, що функція, обернена функції розподілу.

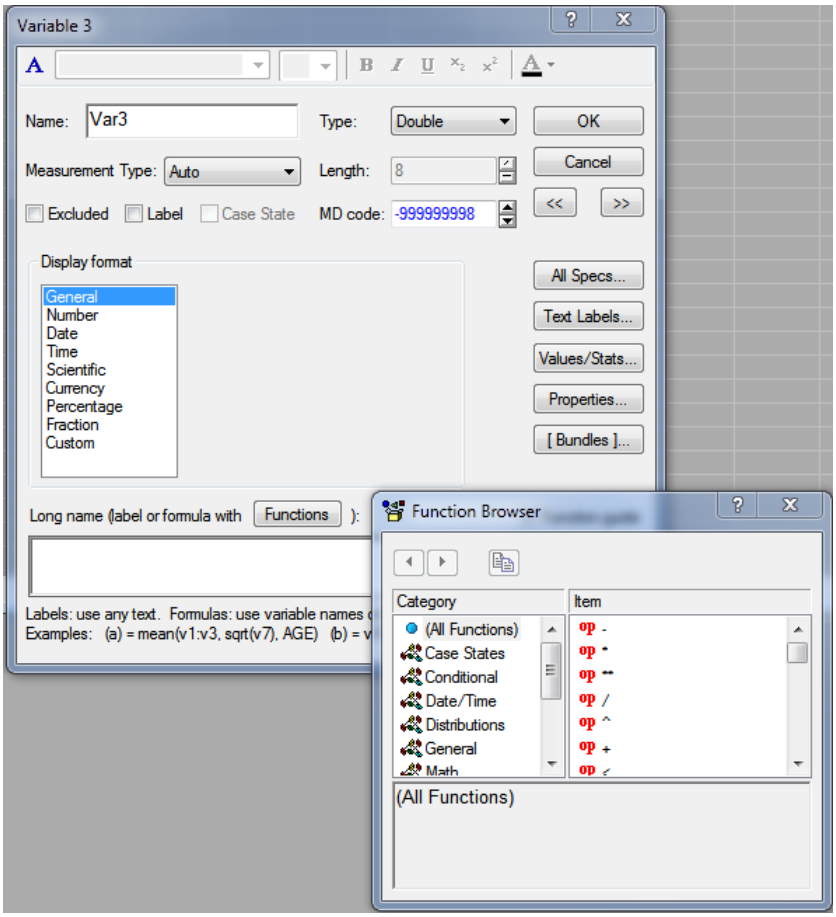


Рис. 7.6. Вибір функції для генерації вибірки

4. Типовий приклад

На основі даних типового прикладу для лабораторної роботи № 3: а) здійснити перевірку на нормальність розподілу для середньої якості ґрунту; б) додати ще дві змінні, розподілені за розподілом Фішера та Стьюдента. Визначити критичні значення критеріїв Фішера, Стьюдента та χ^2 -квадрат; в) визначити з якою ймовірністю випадкова величина попадає в інтервал (120; 150), якщо відомо, що вона розподілена нормально з середнім значенням 135 і середньоквадратичним відхиленням 10,45.

Розв'язування. *Для перевірки гіпотези про нормальність розподілу скористаємося модулем* ***Distribution Fitting.*** *Виберемо у полі* ***Continuous Distribution*** *закон* ***Normal*** *і натиснемо* ***OK.*** *У вікні, що з'явилося, виберемо змінну "Середня якість ґрунту (балів)" та кнопку* ***Plot of Observed and expected distributions****. Гістограма спостережуваних значень змінної "Середня якість ґрунту (балів)" і графік нормального розподілу зображені на рис. 7.7.*

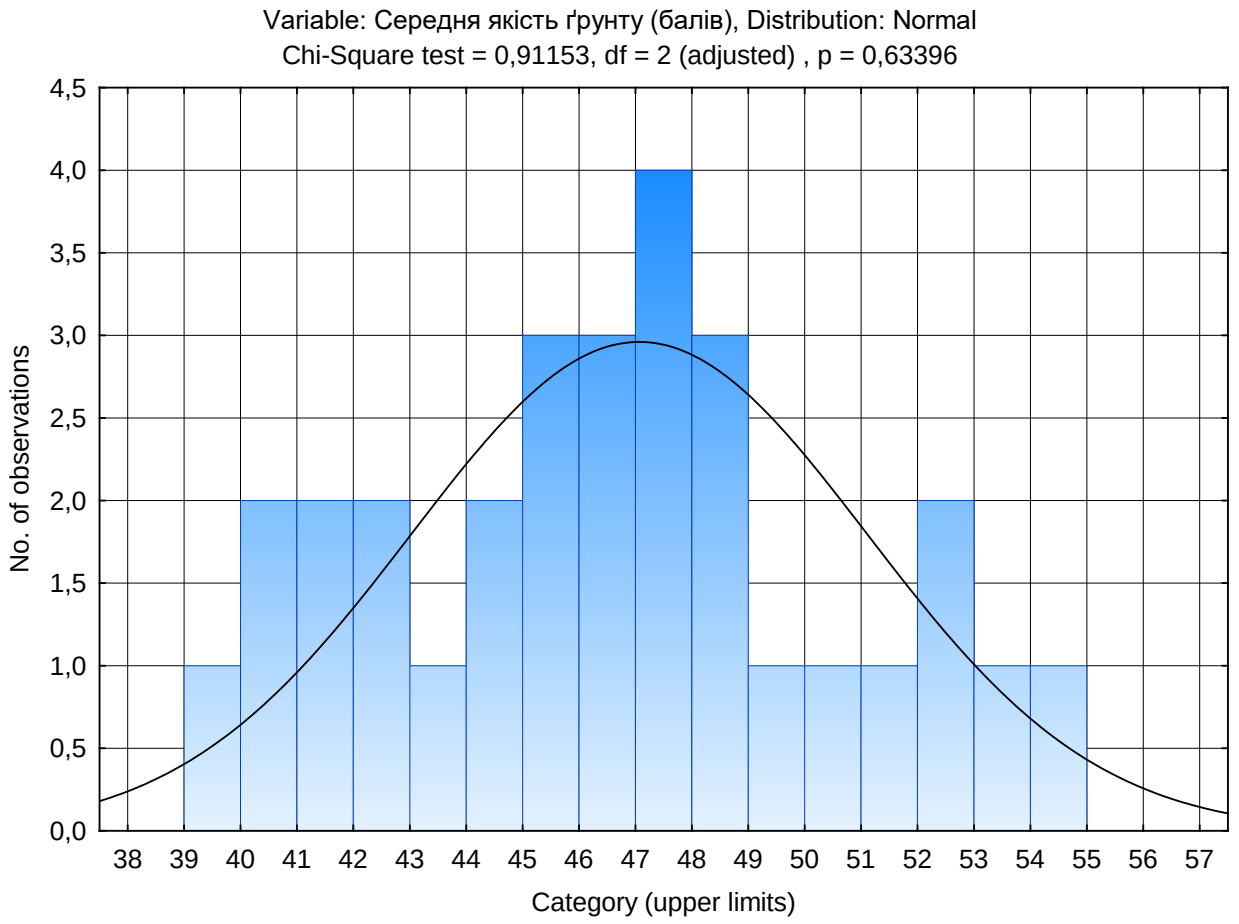


Рис. 7.7. Графік емпіричного і теоретичного розподілів

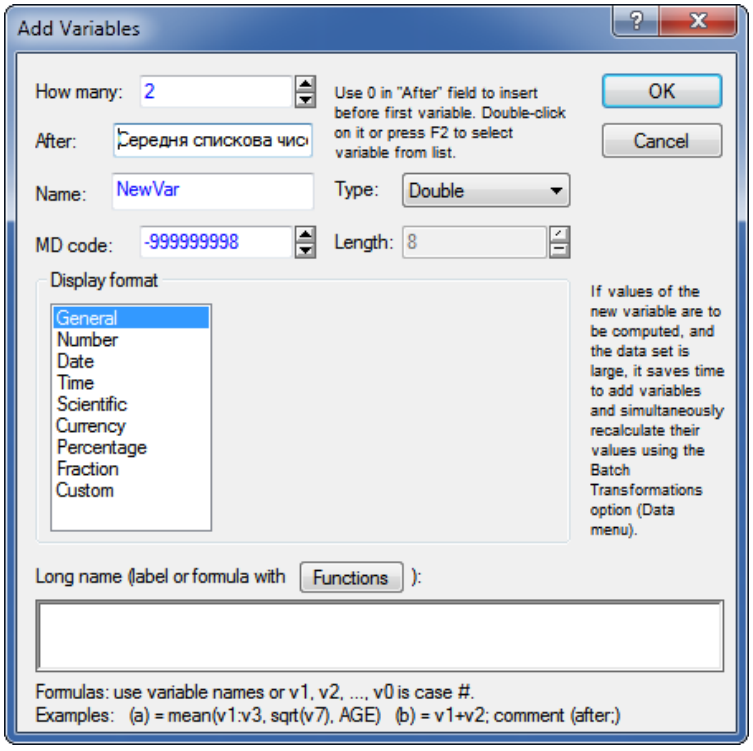


Рис. 7.8. Додавання змінних

На графіку продемонстровано, що досліджувана змінна відповідає нормальному закону розподілу. Якщо ж досліджувана змінна не відповідає нормальному закону підбирають інший закон розподілу у діалоговому вікні ***Distribution Fitting.*** *Також можна змінити початкові дані, виходячи з різниці між*

спостережуваною й очікуваною частотами, що подані в таблиці ***Observed and expected distributions.***

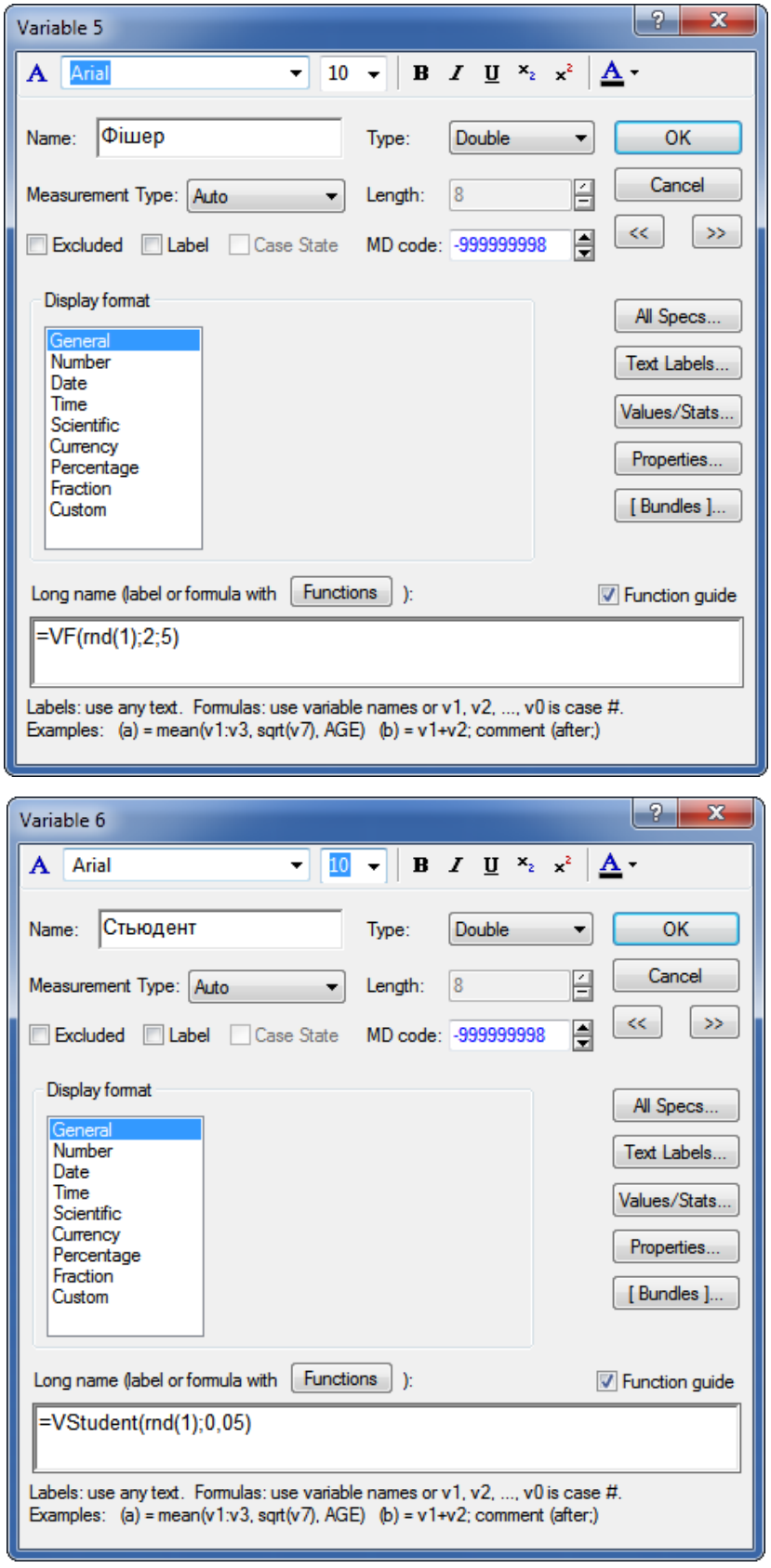


Рис. 7.9. Генерація даних

Щоб додати змінні з розподілом Фішера та Стьюдента, спочатку додамо змінні у таблицю, як це описано в лабораторній роботі №1 (рис. 7.8).

Потім у специфікації змінних введемо формули генерування даних (рис. 7.9).

Щоб визначити критичні значення критеріїв Фішера, Стьюдента та хі-квадрат, скористаємося

модулем ***Probability calculator*** *(рис. 7.10). Уведемо значення p – рівень значущості, df1 та df2 – ступені вільності.*

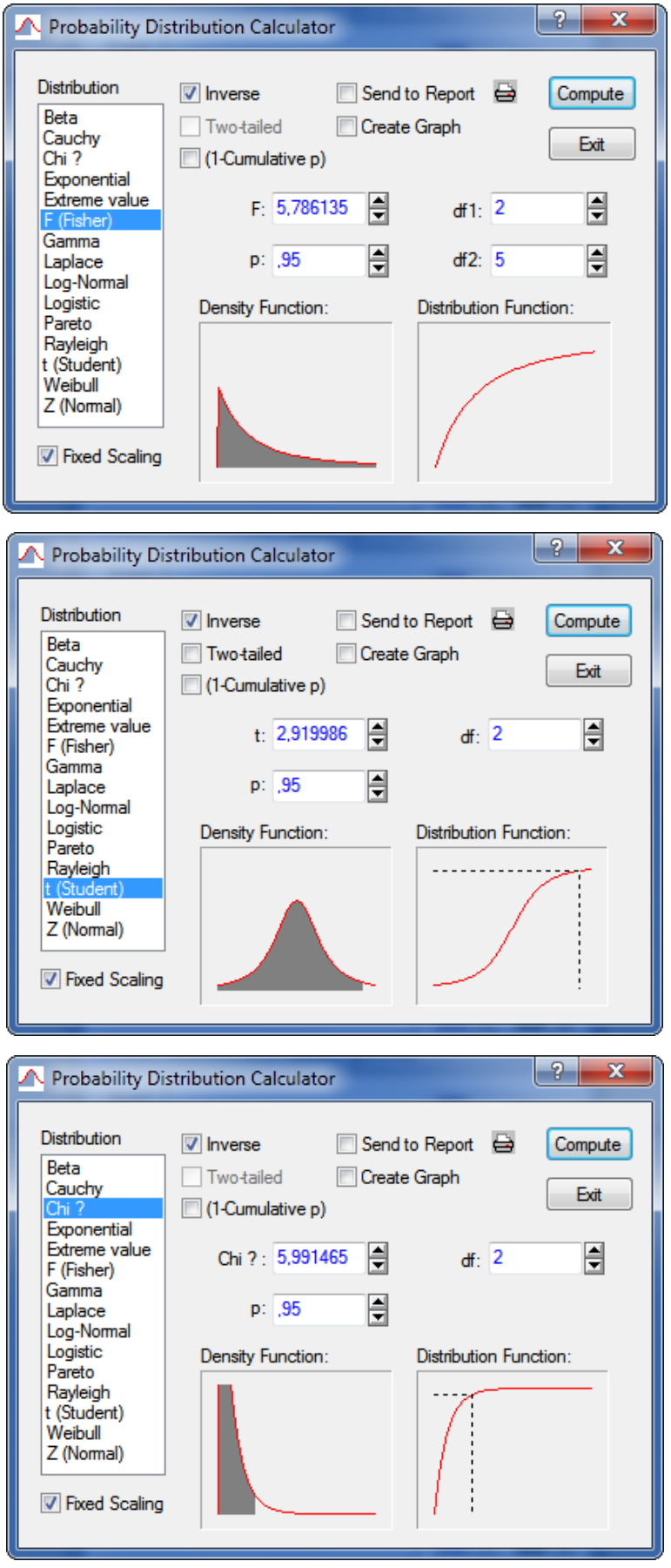


Рис. 7.10. Визначення критичних значень

Щоб визначити ймовірність попадання величини в інтервал використаємо ***Probability calculator.*** *Задамо значення нижньої межі (рис. 7.11) і натиснемо* ***Compute.*** *Ймовірність становитиме 0,075586. Аналогічно розрахуємо ймовірність набуття величиною верхньої межі інтервалу. Дана ймовірність становитиме 0,924414. Отже,* $P(120 < x < 150) = 0.924414 - 0.075586 = 0.848828 \approx 0.84$.

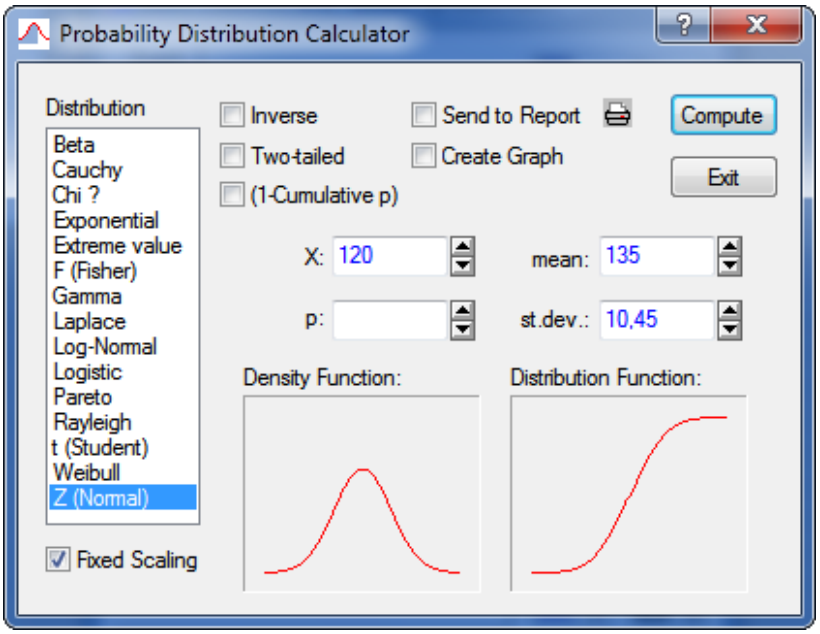


Рис. 7.11. Визначення ймовірності

Завдання для самостійної роботи

7.1. Згенерувати вибірку згідно з даними в таблиці 7.1.

Таблиця 7.1

№ з/п	*Закон*	*Обсяг*	*№ з/п*	*Закон*	*Обсяг*
1	R [0; 2]	50	9	N (1; 4)	60
2	N (2; 0,25)	60	10	E (1)	70
3	E (3)	70	11	R [0; 3]	80
4	R [1; 3]	80	12	N (0; 3)	50
5	N (0; 1)	50	13	E (5)	60
6	E (2)	60	14	R [3; 6]	70
7	R [2; 3]	70	15	N (0; 9)	80
8	N (0; 4)	80	16	E (0,2)	50

У результаті має бути сформований файл з 16 змінними з різними законами розподілу.

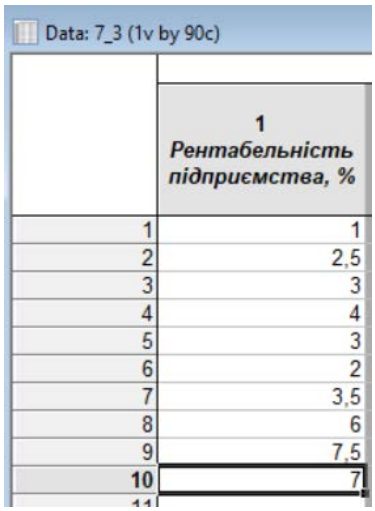
7.2. Знайти ймовірність того, що випадкова величина знаходиться в інтервалі (175; 185), якщо відомо що вона розподілена нормально з параметрами: середнє значення –176,6 і середньоквадратичне відхилення становить 7,63.

7.3. За даними, наведеними в таблиці 7.2, встановити відповідність емпіричного розподілу теоретичному ($\alpha = 0.05$).

Таблиця 7.2

Рентабельність підприємства, %	до 5	5-10	10-15	15-20	20-25	25-30	30 і понад
Кількість підприємств , од.	7	10	15	20	18	13	7

Примітка. Файл для аналізу у STATISTICA формується наступним чином: має бути самостійно задано 7 значень рентабельності в діапазоні 0-5%, 10 значень рентабельності в діапазоні 5%-10% і т.д. У результаті має бути 1 змінна з 90 спостереженнями


Data: 7_3 (1v by 90c)

	1 Рентабельність підприємства, %
1	1
2	2,5
3	3
4	4
5	3
6	2
7	3,5
8	6
9	7,5
10	7

7.4. Побудувати таблиці нормального, χ^2-квадрат, Стьюдента і Фішера розподілів (не менше, ніж 25 значень). Параметри розподілів задайте самостійно.

7.5. Підібрати закон розподілу до завдань лабораторної роботи № 5.

Лабораторна робота № 8
Кореляційний аналіз статистичних даних

1. Основні теоретичні відомості кореляційного аналізу статистичних даних у системі STATISTICA

При вивченні статистичної сукупності часто спостерігається тісний зв'язок між варіаціями досліджуваних ознак (показників) та варіаціями інших ознак (показників), що характеризують дану сукупність. При цьому ознаки, які впливають на зміну інших ознак, називаються ***факторними***, а ознаки, що змінюються під впливом факторних ознак, – ***результативними***. Очевидно, зв'язки між факторними і результативними ознаками є причинно-наслідковими (фактор – причина, результат – наслідок). Крім того, ці зв'язки можуть бути функціональними, статистичними, кореляційними.

Функціональний зв'язок як строга відповідність між змінними x та y може бути однозначним (кожному значенню змінної x відповідає єдине значення змінної y), взаємнооднозначним (кожному значенню змінної x відповідає єдине значення змінної y і навпаки) та багатозначним (принаймі одному значенню змінної x відповідає декілька значень змінної y). Що ж стосується економічних ознак, то між ними може не бути строгої функціональної залежності.

Зовсім інший характер мають так звані статистичні (стохастичні, ймовірнісні) зв'язки, які виникають тоді, коли на змінні x та y впливають багато інших факторів, у тому числі випадкових, що не дозволяє виявити залежності між змінними на основі одиничного випадку. Такі зв'язки можна виявити лише при масовому спостереженні як статистичні закономірності (на основі вивчення особливостей розподілу, поведінки середніх та інших показників). Виявлені у такий спосіб зв'язки і будуть статистичними.

Статистичним (стохастичним) називається такий зв'язок між змінними x та y, коли кожному значенню однієї змінної відповідає певний умовний розподіл іншої змінної.

Кореляційним зв'язком між змінними x, y називається числова функціональна залежність між значеннями однієї з них та умовними середніми значеннями (умовними математичними сподіваннями) іншої.

У даному визначенні кореляційного зв'язку у ролі незалежної змінної виступає ознака-фактор, а залежної змінної – умовна середня величина результативної ознаки. При цьому така кореляція називається ***парною***. Якщо факторних ознак m $(m>1)$, а результативна одна, то кореляція називається ***множинною***. При вивченні множинної кореляції вводиться також поняття ***частинної*** кореляції, під якою розуміється кореляційна залежність між результативним показником y та одним із факторних показників x_i (i – номер одного з m показників) в умовах, коли вплив на них інших факторів (на певному фіксованому рівні) усунений.

Для виявлення кореляційного зв'язку та його характеру система **STATISTICA** пропонує ряд методів.

Розглянемо один з них – ***Correlation matrices (Кореляційні матриці)***.

Для запуску модуля необхідно у вкладці ***Statistics*** у групі ***Base*** в модулі ***Basic Statistics/Tables*** або в меню ***Statistics*** у модулі ***Basic Statistics/Tables*** вибрати команду ***Correlation matrices.*** Відкриється діалогове вікно ***Product-Moment and Partial Correlations (Лінійні та частинні коефіцієнти кореляції)*** (рис. 8.1).

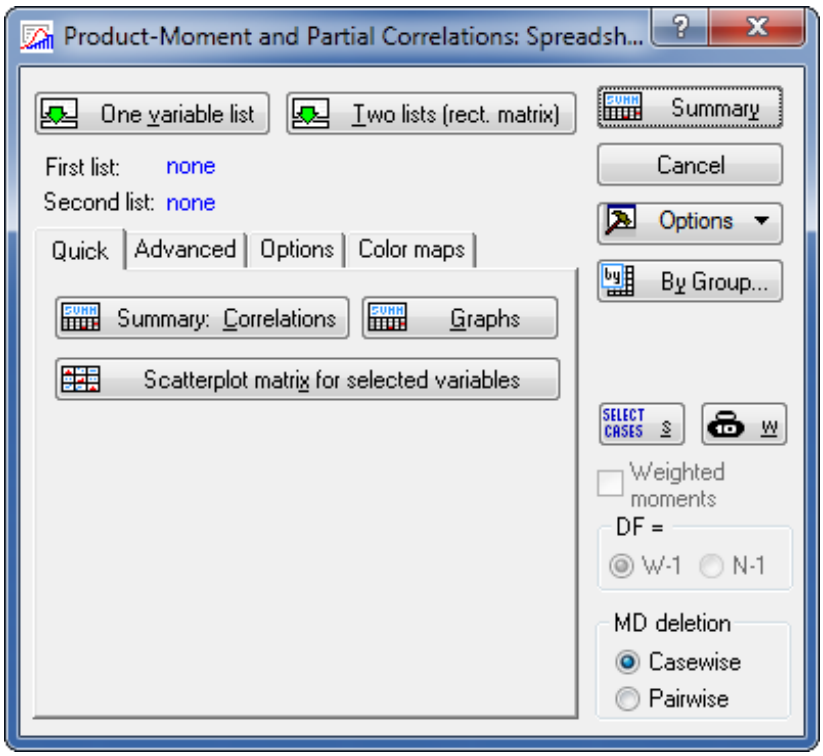


Рис. 8.1. Діалогове вікно ***Product-Moment and Partial Correlations***

Існує два шляхи формування матриць кореляцій: ***One variable list (Один список змінних)*** та ***Two lists (rect. matrix) (Два списки змінних (прямокутна матриця))*** (рис. 8.1).

Вибравши кнопку ***One variable list*** діалогового вікна (рис. 8.1), отримаємо перелік змінних, які потрібно вибрати для оцінки кореляційного зв'язку між ними. Кнопка ***Two lists (rect. matrix)*** дозволяє вибрати змінні у двох змінних, зазначивши імена змінних для стовпчиків і рядків кореляційної матриці (рис. 8.2). Після вибору змінних і натискання кнопки ***Summary: Correlations (Результати: Коефіцієнти кореляції)*** з'являється матриця зі значеннями коефіцієнтів кореляції.

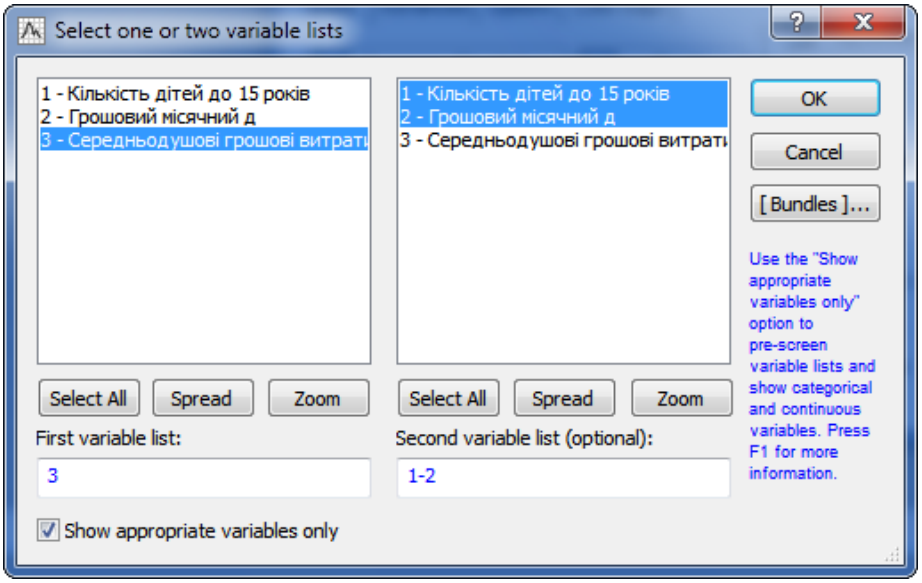


Рис. 8.2. Вибір змінних для побудови прямокутної матриці кореляцій

Крім кнопки ***Summary: Correlations***, натискання якої дозволяє побудувати матриці коефіцієнтів кореляції, на вкладці ***Quick*** діалогового вікна (рис. 8.1) ще розташовані кнопки додаткового аналізу:

- ***Graphs (Графіки)***;
- ***Scatterplot matrix for selected variables (Графіки розсіювання для обраних змінних)***.

Кнопка ***Graphs*** будує гістограми емпіричного та теоретичного розподілів для кожної змінної та виводить основні статистичні показники (верхня частина рис. 8.3), діаграму розсіювання з довірчими інтервалами та рівняння регресії.

Scatterplot matrix for selected variables будує гістограми для кожної змінної окремо та діаграми розсіювання (рис. 8.4).

2. Додаткові можливості здійснення кореляційного аналізу в системі STATISTICA

Вкладка ***Advanced*** діалогового вікна – вікно ***Product-Moment and Partial Correlations*** (рис. 8.1) використовується для додаткового кореляційного аналізу і має аналогічні кнопки як у вкладці ***Quick***, а також і додаткові, зокрема:

- ***Matrix 1 (Матриця 1)*** – побудова матриці кореляції для одного списку змінної в окремому вікні;
- ***Partial correlations (Частинна кореляція)*** – побудова матриці частинних коефіцієнтів кореляції;
- ***Matrix 2 (Матриця 2)*** – побудова матриці частинних коефіцієнтів кореляції в окремому вікні.

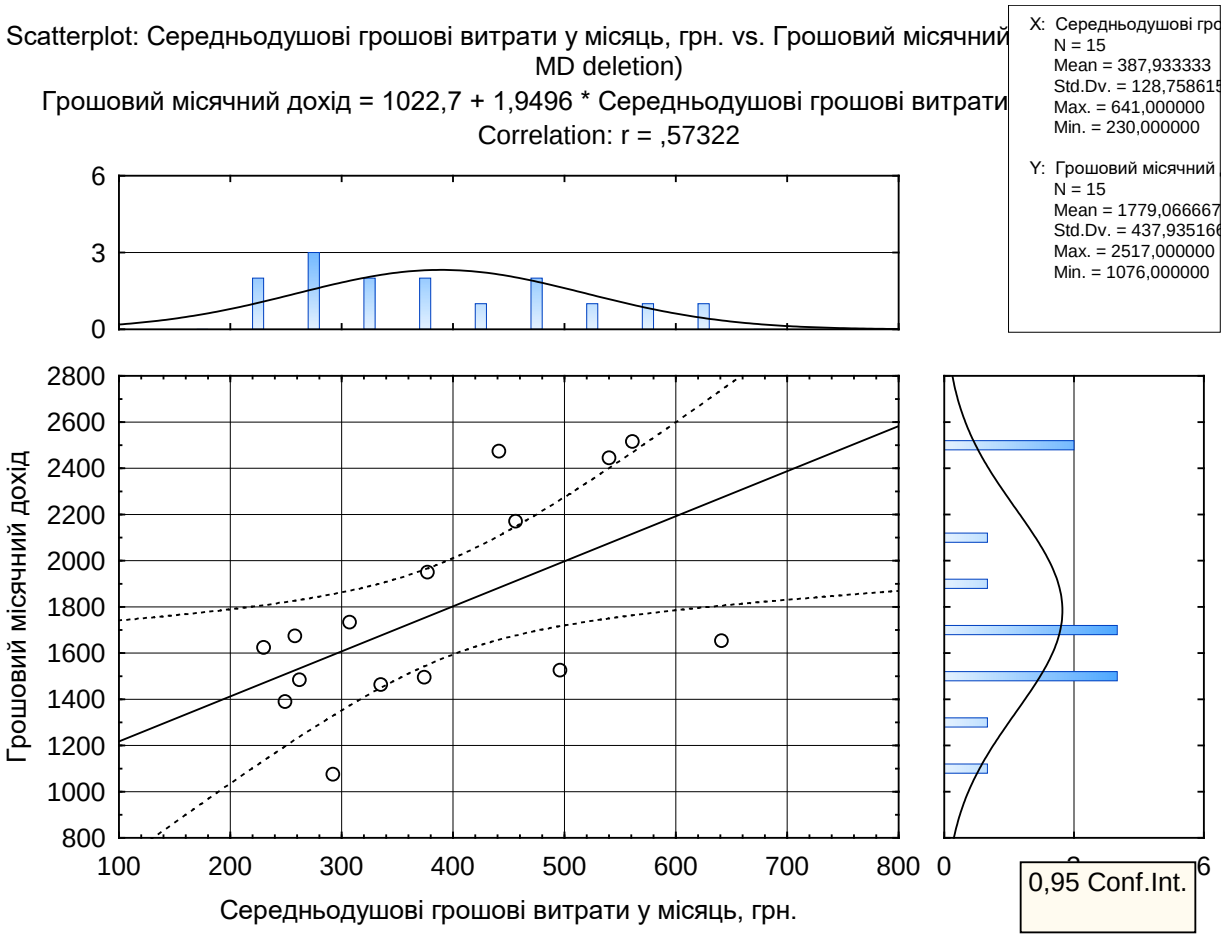


Рис. 8.3. Результат графічного аналізу ***Graphs***

Система **STATISTICA** у вкладці ***Advanced*** основного діалогового вікна кореляційного аналізу пропонує цілий комплекс графічних засобів аналізу статистичних даних. У модулі кореляційного аналізу можна використати такі типи графіків:

- ***2D scatterplots*** – двовимірні діаграми розсіювання. Якщо задати ***with casenames (з іменами спостережень)***, то побудується аналогічний графік з пойменованими точками;
- ***3D scatterplots –*** тривимірні діаграми розсіювання (можна подати і з іменами спостережень);
- ***Scatterplot matrix*** – аналогічно вкладки ***Quick***;
- ***Categ. scatterplots*** – категоризовані діаграми розсіювання, будуються у тих випадках, якщо змінна має спостереження, що належать до різних груп, наприклад типи підприємств, територіальні одиниці тощо. Після вибору змінних з'являється вікно вибору групувальної змінної. Одночасно можна вибрати не більше, ніж дві змінні, за якими відбуватиметься категоризація;
- ***Surface plots*** – графіки поверхонь;
- ***3D histogram –*** тривимірна гістограма.

Вкладка ***Options*** діалогового вікна (рис. 8.1) містить елементи задання формату виведення результатів у полі ***Display format for correlation matrices (Формат показу для мат риць кореляції)*** (рис. 8.5).

У даному полі доступні такі формати:

- ***Display simple matrix (highlight p's) (Показати просту матрицю (p підсвічувати))*** – проста матриця значень коефіцієнтів кореляції. Значимі значення коефіцієнтів кореляції виділяються червоним кольором.
- ***Display r, p-levels, and N's (Показати r, p-рівень і N's)*** – матриця із значеннями коефіцієнтів кореляції, рівня значущості *p* і обсяг вибірки.
- ***Display detailed table of results (Показати детальну таблицю результатів)*** – матриця із загальними середніми значеннями, парними середньоквадратичними відхиленнями та інша регресійна статистика.

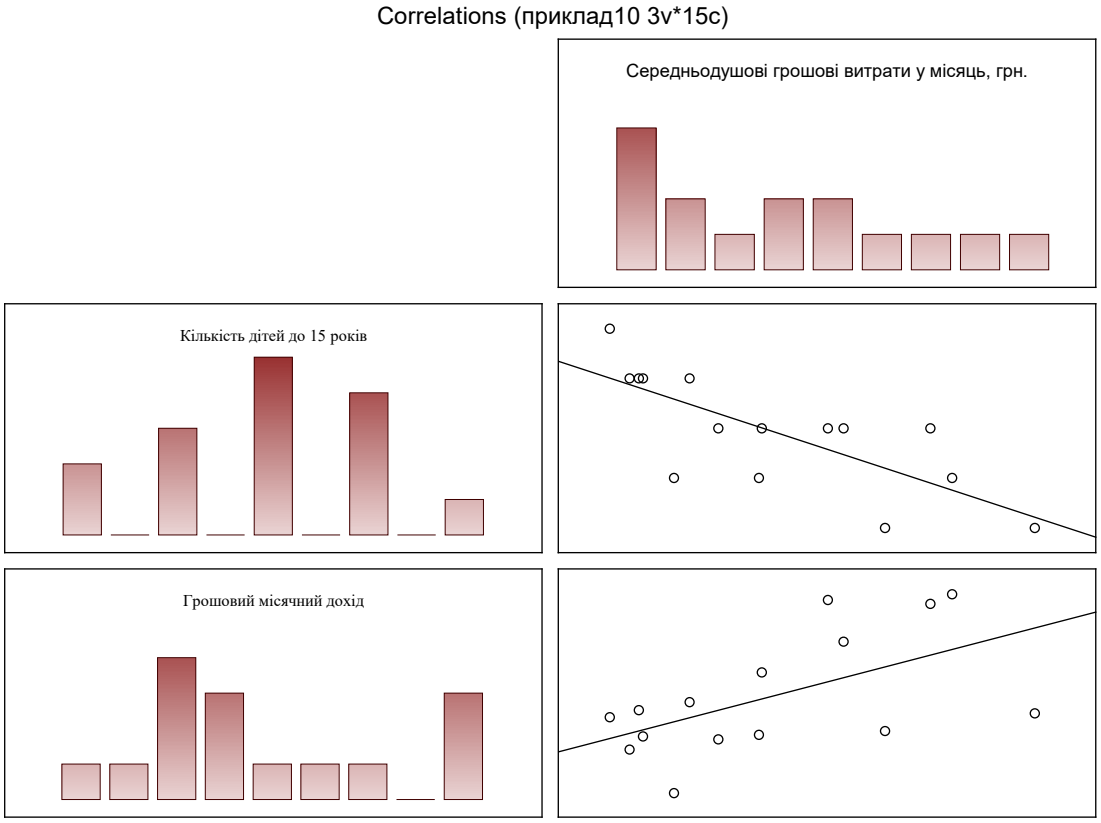


Рис. 8.4. Результат побудови ***Scatterplot matrix for selected variables***

Display long variable names (Відображення довгих імен змінних) та ***Extended precision***

calculations (Подвійна точність розрахунків) і ***MD deletion*** (***Видалення пропущених даних***) мають зміст аналогічний, як і в модулі описової статистики ***(Descriptive statistics)***.

У полі ***p-value for highlighting (p-рівень для підсвічування)*** задається значення для рівня значущості.

Якщо встановити позначку ***Include means and std. devs. in square matrices (Включати середнє та середньоквадратичне відхилення у квадратну матрицю)***, то матриці кореляції міститимуть середнє значення та середньоквадратичне відхилення.

На вкладці ***Color maps (Кольорові карти)*** діалогового вікна (рис. 8.1) задані опції для створення кольорових карт, де величина або статистична значущість коефіцієнта кореляції в кореляційній матриці зазначається у відповідному кольорі комірці, що відповідає певному інтервалу, який характеризує напрямок і тісноту зв'язку.

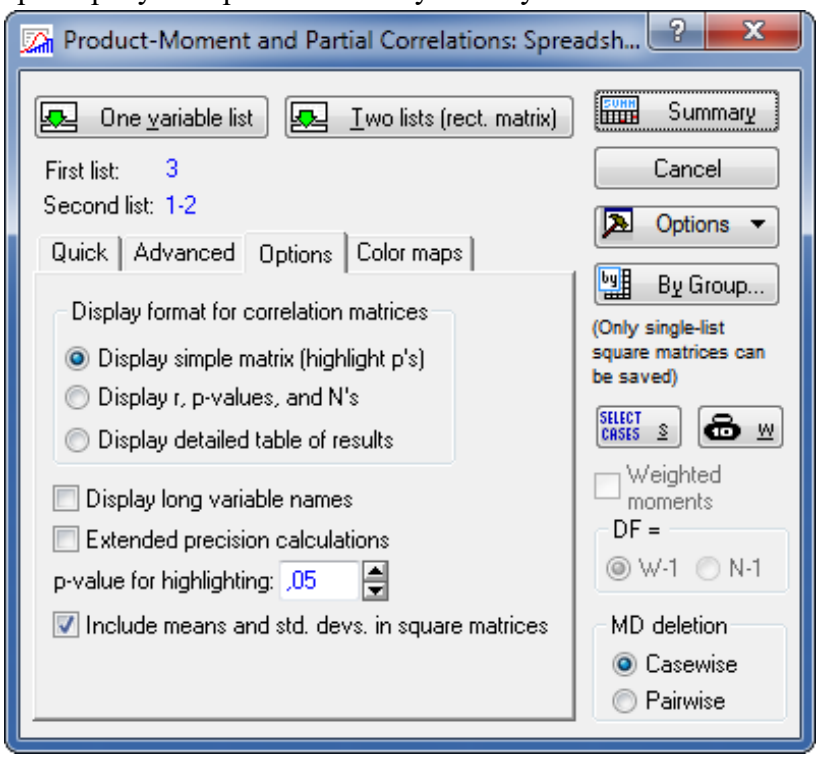


Рис. 8.5. Редагування формату виведення матриці кореляції

У полі ***Display format for correlation matrices*** можна задати опції подання кольорових карт: ***Simple matrix (r-values) (Проста матриця (коефіцієнти кореляції))***, ***Matrix of absolute r values (Матриця абсолютних значень коефіцієнтів кореляції)*** та ***Matrix of p values (Матриця рівнів значущості).*** Якщо натиснути кнопку ***Color maps,*** буде побудована кольорова матриця (рис. 8.6), залежно від указаного типу.

Color map of p-values for correlations (приклад10)
Marked correlations are significant at p < ,05000
N=15 (Casewise deletion of missing data)

Variable	Кількість дітей до 15 років	Грошовий місячний дохід	Середньодушові грошові витрати у місяць, грн.
p<=	0,001	0,010	0,025	0,050	0,100	0,150	0,200	0,350	0,500	1	
Кількість дітей до 15 років		0,947	0,001
Грошовий місячний дохід	0,947		0,025
Середньодушові грошові витрати у місяць, грн.	0,001	0,025	

Рис. 8.6. Кольорова матриця кореляції

3. Типовий приклад

На основі вибіркової сукупності даних (табл. 8.1) визначити парні та частинні коефіцієнти кореляції.

Таблиця 8.1

Номер спостереження	*Виробничі фонди (млн грн)* (x_1)	*Трудові ресурси (млн грн)* (x_2)	*Випуск продукції (млн грн)* (y)
1	*10*	*12*	*20*
2	*15*	*10*	*35*
3	*20*	*9*	*30*
4	*25*	*11*	*45*
5	*40*	*12*	*60*
6	*37*	*11*	*69*
7	*43*	*13*	*75*
8	*35*	*15*	*90*
9	*38*	*16*	*105*
10	*55*	*18*	*110*

Розв'язування. *Скористаємося модулем* ***Correlation matrices.*** *Виберемо змінні (рис. 8.7) та побудуємо матрицю парних коефіцієнтів кореляції, використовуючи кнопку* ***Summary: Correlations.*** *Оскільки на вкладці* ***Options*** *вказано* ***Include means and std. devs. in square matrices****, у таблиці подані значення середньої величина та середньоквадратичне відхилення змінних (рис. 8.8). Для візуалізації залежності використаємо* ***2D scatterplots*** *(рис. 8.9). На графіку проілюстровано також рівняння регресії.*

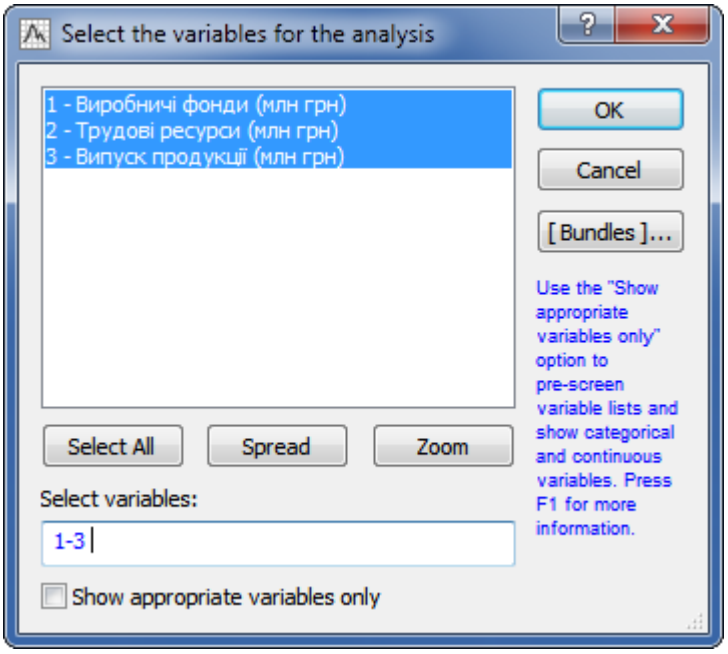


Рис. 8.7. Вибір змінних

Correlations (Spreadsheet8)
Marked correlations are significant at p < ,05000
N=10 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	Виробничі фонди (млн грн)	Трудові ресурси (млн грн)	Випуск продукції (млн грн)
Виробничі фонди (млн грн)	31,80000	13,94274	1,000000	0,719091	0,884966
Трудові ресурси (млн грн)	12,70000	2,83039	0,719091	1,000000	0,877043
Випуск продукції (млн грн)	63,90000	31,49762	0,884966	0,877043	1,000000

Рис. 8.8. Матриця кореляції

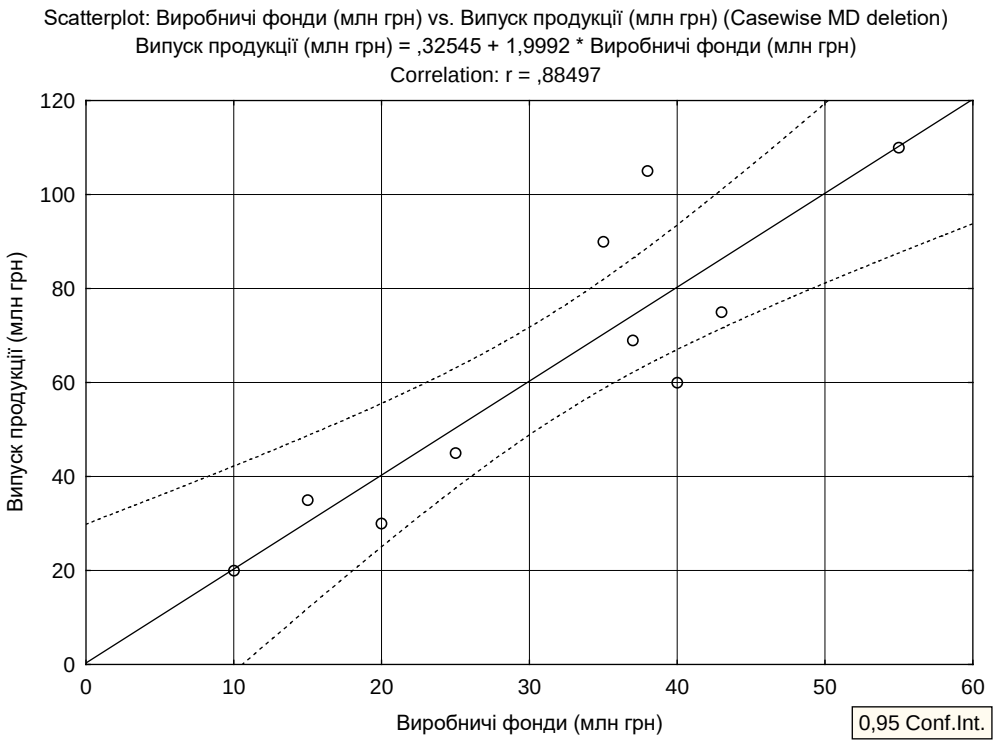


Рис. 8.9. Графічний аналіз тісноти зв'язку

Для визначення частинних коефіцієнтів кореляції натиснемо кнопку ***Partial correlations*** *у вкладці* ***Advanced*** *діалогового вікна вікно* ***Product-Moment and Partial Correlations.*** *Результат поданий на рис. 8.10.*

Partial Correlations (приклад10)
Controlling for: Кількість дітей до 15 років
Marked correlations are significant at p < ,05000
N=15 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	Кількість дітей до 15 років	Грошовий місячний дохід	Середньодушові грошові витрати у місяць, грн.
Кількість дітей до 15 років	1,933	1,1629	1,000000		
Грошовий місячний дохід	1779,067	437,9352		1,000000	0,848622
Середньодушові грошові витрати у місяць, грн.	387,933	128,7586		0,848622	1,000000

Рис. 8.10. Матриця частинних коефіцієнтів кореляції

Завдання для самостійної роботи

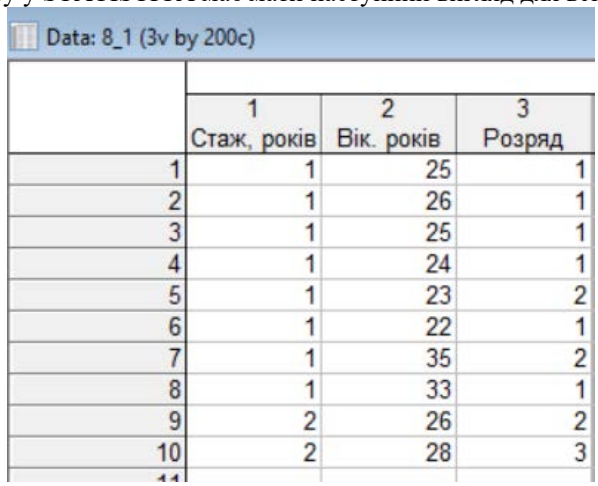
8.1. Здійснити кореляційний аналіз на основі даних про працівників певного підприємства (табл. 8.2). (Електронну таблицю даних формують 4 студенти, а потім зливають в один файл і виконують подальший аналіз).

Таблиця 8.2

№ з/п	*Стаж, років*	*Вік. років*	*Розряд*	*№ з/п*	*Стаж, років*	*Вік, років*	*Розряд*	*№ з/п*	*Стаж, років*	*Вік, років*	*Розряд*
1	1	25	1	68	4	41	4	135	6	33	4
2	1	26	1	69	4	41	4	136	6	35	4
3	1	25	1	70	4	41	4	137	6	36	4
4	1	24	1	71	4	41	4	138	6	35	4
5	1	23	2	72	4	41	4	139	6	50	5
6	1	22	1	73	5	42	4	140	6	52	5
7	1	35	2	74	5	42	4	141	6	50	5
8	1	33	1	75	5	42	4	142	6	49	5
9	2	26	2	76	5	42	4	143	6	48	5
10	2	28	3	77	5	42	4	144	6	49	5
11	2	30	2	78	5	42	4	145	6	49	5
12	2	33	3	79	5	45	4	146	6	49	5
13	2	34	2	80	5	43	4	147	6	47	5
14	2	50	3	81	5	46	4	148	7	46	5
15	2	42	3	82	5	45	4	149	7	53	5
16	2	40	3	83	5	45	4	150	7	53	5
17	2	23	3	84	5	42	4	151	7	53	5
18	2	30	2	85	5	41	4	152	7	53	5
19	2	32	3	86	5	40	4	153	7	53	6
20	2	29	2	87	5	40	4	154	7	55	5
21	3	30	3	88	5	40	4	155	7	52	6
22	3	36	2	89	5	40	4	156	7	52	5
23	3	38	3	90	5	36	3	157	7	51	6
24	3	37	2	91	5	39	3	158	7	51	5
25	3	35	3	92	5	34	3	159	7	51	6
26	3	39	2	93	5	35	3	160	7	50	5
27	3	40	3	94	5	35	4	161	7	50	6
28	3	41	4	95	5	35	3	162	7	52	5
29	3	44	4	96	5	36	4	163	7	52	4
30	3	29	4	97	5	38	4	164	7	52	5
31	3	30	4	98	5	37	4	165	7	52	4
32	3	35	4	99	5	39	3	166	7	52	5
33	3	36	3	100	5	42	4	167	7	52	4
34	3	37	3	101	5	43	3	168	7	52	5
35	3	38	2	102	5	44	4	169	7	49	4
36	3	35	3	103	5	44	3	170	7	48	5
37	3	36	2	104	5	44	4	171	7	47	4
38	3	37	3	105	5	44	3	172	7	48	5
39	3	38	4	106	5	44	4	173	8	49	4
40	3	35	4	107	5	44	3	174	8	52	5
41	4	36	4	108	5	44	4	175	8	52	4
42	4	37	4	109	5	44	3	176	8	52	5
43	4	38	4	110	5	44	4	177	8	52	4
44	4	39	4	111	5	44	4	178	8	52	5

№ з/п	*Стаж, років*	*Вік. років*	*Розряд*	*№ з/п*	*Стаж, років*	*Вік, років*	*Розряд*	*№ з/п*	*Стаж, років*	*Вік, років*	*Розряд*
45	4	39	4	112	5	44	4	179	8	52	4
46	4	39	3	113	5	44	4	180	8	52	5
47	4	39	3	114	5	44	4	181	8	53	6
48	4	39	4	115	5	44	4	182	8	53	5
49	4	40	4	116	6	44	4	183	8	54	6
50	4	41	4	117	6	44	4	184	8	54	5
51	4	42	4	118	6	45	4	185	8	54	6
52	4	41	4	119	6	46	4	186	9	54	5
53	4	42	4	120	6	48	4	187	9	55	6
54	4	41	5	121	6	49	4	188	9	55	5
55	4	42	4	122	6	48	4	189	9	55	6
56	4	41	4	123	6	47	4	190	9	56	5
57	4	42	4	124	6	48	4	191	9	58	6
58	4	41	4	125	6	49	4	192	9	54	6
59	4	28	2	126	6	47	4	193	9	56	6
60	4	40	4	127	6	52	4	194	9	52	6
61	4	39	4	128	6	50	5	195	9	53	6
62	4	36	3	129	6	51	5	196	10	54	6
63	4	35	3	130	6	52	4	197	10	59	6
64	4	36	3	131	6	51	5	198	10	58	6
65	4	39	3	132	6	52	4	199	10	55	6
66	4	40	3	133	б	51	5	200	10	55	6
67	4	41	4	134	6	48	5				

Примітка. Файл для аналізу у STATISTICA має мати наступний вигляд для всіх спостережень.


Data: 8_1 (3v by 200c)

	1 Стаж, років	2 Вік. років	3 Розряд
1	1	25	1
2	1	26	1
3	1	25	1
4	1	24	1
5	1	23	2
6	1	22	1
7	1	35	2
8	1	33	1
9	2	26	2
10	2	28	3

8.2. У таблиці 8.3 наведено дані про дохід і споживчі витрати домогосподарств. Визначити напрямок та тісноту зв'язку між доходом домогосподарств і споживчими витратами. Для здійснення кореляційного аналізу використати всі можливості вкладки ***Advanced*** модуля ***Correlation matrices***.

Таблиця 8.3

Номер домогосподарства	Дохід (грн.)	Споживчі витрати (грн.)
1	1000	600
2	1100	610
3	1150	620
4	1300	700
5	1400	750
6	1450	750

Номер домогосподарства	Дохід (грн.)	Споживчі витрати (грн.)
7	1550	780
8	1600	800
9	1800	900
10	2000	950

8.3. Визначити взаємозв'язок між вмістом фтору у питній воді, ураженістю зубів карієсом і флюорозом (табл. 8.4). Результати подати у вигляді детальної кореляційної таблиці. Використати подвійну точність, використовуючи ***Extended precision calculations*** на вкладці ***Options*** діалогового вікна ***Product-Moment and Partial Correlations***, обчислити частинні коефіцієнти кореляції та пояснити їх зміст.

Таблиця 8.4

Вміст фтору у воді (мг/л)	*Ураженість зубів флюорозом (%)*	*Ураженість зубів карієсом (%)*
0,3	0	75
0,8	18	43
1,3	25	38
1,8	35	25
2,5	45	23
4,0	85	13

8.4. Результати оцінювання групи країн за ступенем етнічного різноманіття та соціальної напруженості подано у таблиці 8.5. Визначити коефіцієнти кореляції, побудувати кольорову карту і здійснити графічний аналіз.

Таблиця 8.5

Країна	*Етнічне різноманіття, бали*	*Ступінь соціальної напруженості, бали*
А.	93	126
Б.	85	125
В.	82	478
Г.	80	502
Д.	75	100
Е.	60	7
Ж.	45	78
З.	43	75
И.	42	70
К.	39	126
Л.	29	7
М.	25	52
Н.	15	48
О.	5	48
П.	0	33

8.5 У результаті маркетингового дослідження глядачів кінотеатру отримані дані (табл. 8.6). Оцінити тісноту зв'язку між періодом транслювання реклами та її запам'ятовуванням. Провести детальний та графічний аналіз.

Таблиця 8.6

Реклама	*Запам'ятовування реклами*		
	Добре запам'яталася	*Залишилася згадка*	*Не запам'яталася*
Перед фільмом	40	20	10
Під час фільму	30	16	24
Після фільму	6	24	30
Разом	76	60	64

Лабораторна робота № 9
Регресійний аналіз статистичних даних

1. Основні теоретичні відомості кореляційно-регресійного аналізу в системі STATISTICA

Кореляційно-регресійний аналіз – всебічне дослідження кореляційних зв'язків, яке зводиться до розв'язання трьох основних задач:

- установлення (виявлення) факту наявності кореляційного зв'язку між досліджуваними ознаками;
- вимірювання тісноти зв'язку за допомогою спеціальних коефіцієнтів (кореляційний аналіз);
- побудова й оцінювання рівняння регресії – математичної моделі функції однієї чи багатьох змінних, у якій аргументи є факторами, а залежна змінна – середня величина результативної ознаки (регресійний аналіз).

У системі **STATISTICA** кореляційного-регресійній аналіз проводиться в модулях: ***Multiple regression (Множинна регресія)*** і ***Nonlinear Estimation (Нелінійне оцінювання).*** Загальне призначення вказаних модулів – це побудува регресійної моделі, що описує кореляційні зв'язки між досліджуваними змінними, оцінка значущості побудованої регресійної моделі. Модуль ***Multiple regression*** використовується для побудови лінійних регресійних моделей, модуль ***Nonlinear Estimation*** – для побудови нелінійних регресійних моделей.

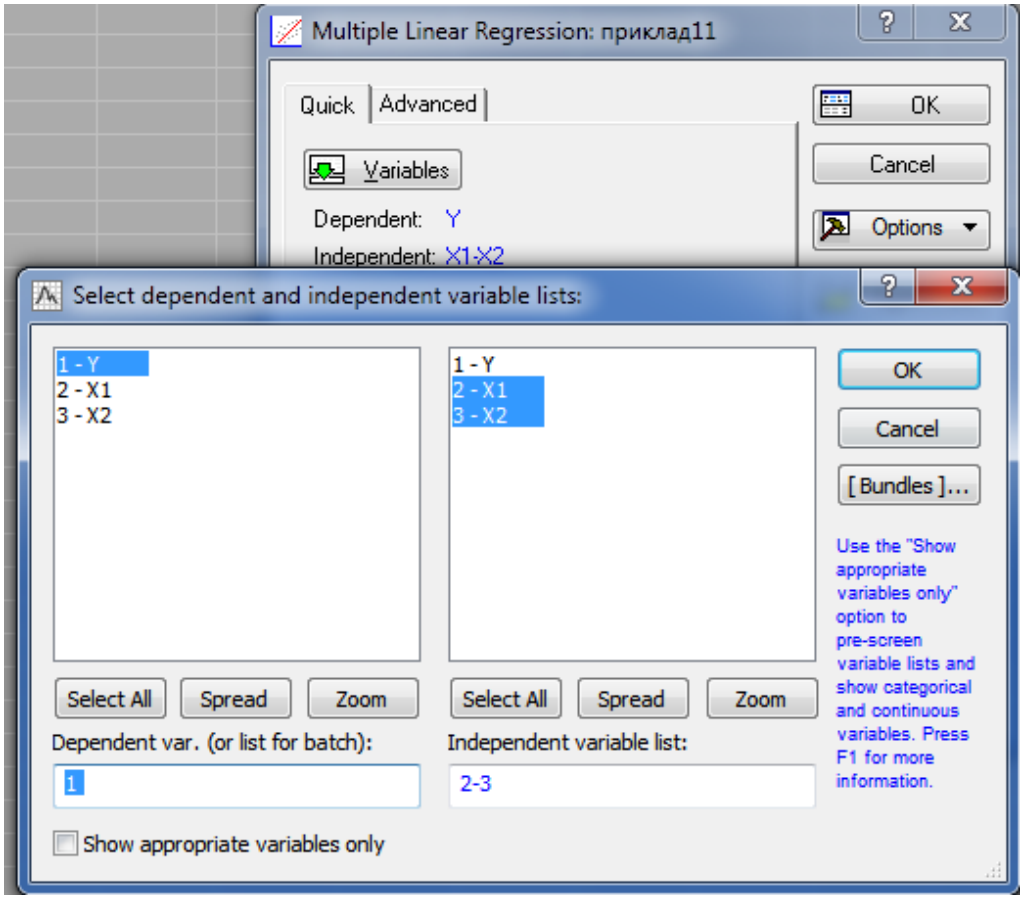


Рис. 9.1. Вибір змінних для побудови регресійної моделі

Розглянемо модуль ***Multiple regression.*** Для запуску модуля необхідно у вкладці ***Statistics*** у групі ***Base*** або в меню ***Statistics*** вибрати команду ***Multiple regression.*** Вибір змінних для побудови регресійної моделі здійснюється за допомогою кнопки ***Variables*** (рис. 9.1). Натиснувши кнопку ***Variables***, відкриється діалогове вікно ***Select dependent and independent variable list (Вибрати зі списку залежних і незалежних змінних)*** (рис. 9.1). Залежна змінна

вибирається в лівій частині вікна, незалежні змінні – у правій частині (рис. 9.1). Номери змінних запишуться відповідно в ***Dependent var. (or list for batch) (Список залежних змінних)*** і ***Independent variable list (Список незалежних змінних)***.

Після натискання ***OK*** система знову повернеться в стартове діалогове вікно ***Multiple Linear Regression (Множинна лінійна регресія)***. У діалоговому вікні ***Multiple Linear Regression*** кнопка ***OK*** виведе результати регресійного аналізу (рис. 9.2).

Опишемо вікно результатів кореляційно-регресійного аналізу (рис. 9.2). Верхня частина вікна результатів – інформаційна. У першій частині міститься основна інформація про результати оцінювання, у другій – висвічуються значущі стандартизовані регресійні коефіцієнти. Внизу вікна знаходяться функціональні кнопки, які дозволяють переглянути результати аналізу.

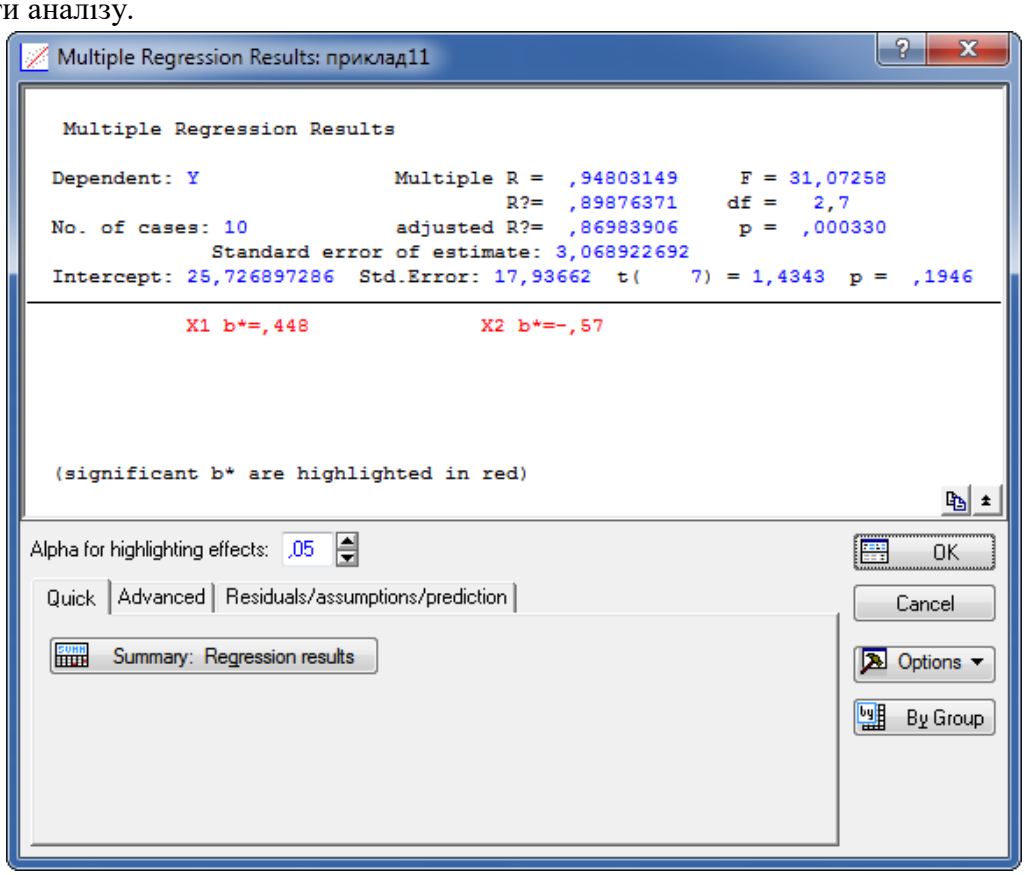


Рис. 9.2. Вікно результатів кореляційно-регресійного аналізу

В інформаційній частині містяться короткі відомості про результати аналізу, а саме:

- ***Dependent (Ім'я залежної змінної)***;
- ***No. of cases (Число спостережень)***;
- ***Multiple R (Коефіцієнт множинної кореляції)***;
- ***R? (Коефіцієнт детермінації)***;
- ***adjusted R? (Скоригований коефіцієнт детермінації)***;
- ***Standard error of estimate (Стандартна помилка оцінки)***;
- ***Intercept (Оцінка вільного члена регресії)***;
- ***Std. Error (Стандартна помилка оцінки вільного члена)***;
- ***t, p (Значення t-критерію і рівень значущості p)***;
- ***F, df, p (Значення F-критерію, число ступенів вільності і рівень значущості p)***.

Під лінією подані стандартизовані оцінки коефіцієнтів регресії.

Якщо вибрати кнопку ***Summary: Regression results (Результати: Регресія)*** на рис. 9.2, то

система виведе результати у вигляді таблиці (рис. 9.3).

Regression Summary for Dependent Variable: Y (Spreadsheet1)
R= ,94803149 R?= ,89876371 Adjusted R?= ,86983906
F(2,7)=31,073 p<,00033 Std.Error of estimate: 3,0689

N=10	b*	Std.Err. of b*	b	Std.Err. of b	t(7)	p-value
Intercept			25,72690	17,93663	1,43432	0,194605
X1	0,447815	0,175840	1,12819	0,44300	2,54672	0,038286
X2	-0,570488	0,175840	-3,16480	0,97548	-3,24436	0,014167

Рис. 9.3. Таблиця результатів регресійного аналізу

Таблиця містить стандартизовані (b*) і звичайні (b) оцінки коефіцієнтів регресії, їх стандартні помилки, *t*-критерії (в дужках указано ступені вільності) і рівні значущості. Коефіцієнти b* оцінюються за стандартизованими даними, мають вибіркове середнє значення 0 і середньоквадратичне відхилення 1.

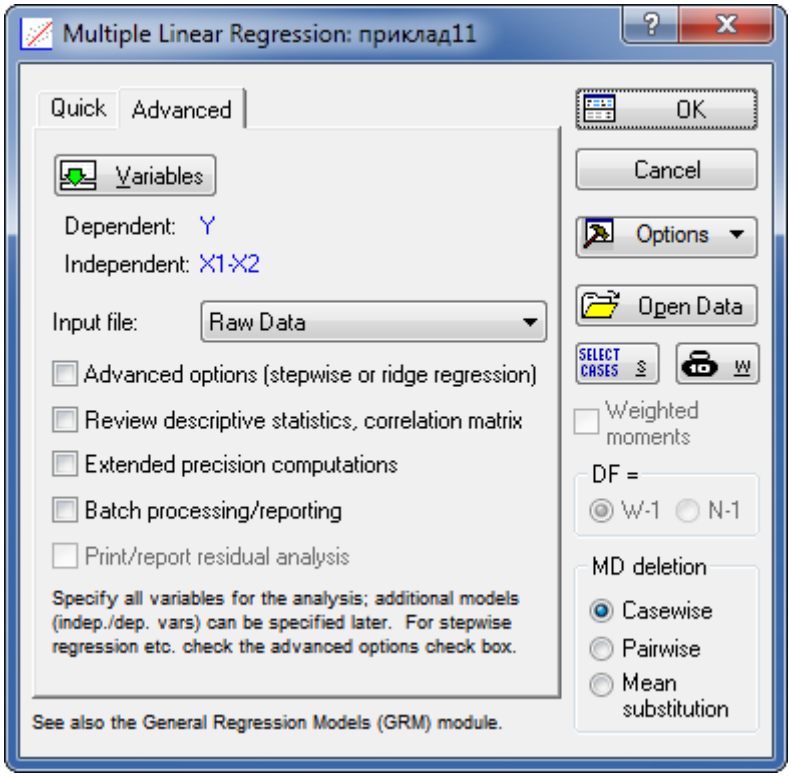


Рис. 9.4. Вкладка ***Advanced*** діалогового вікна ***Multiple Linear Regression***

Для задання додаткових можливостей регресійного аналізу треба вибрати вкладку ***Advanced*** діалогового вікна ***Multiple Linear Regression***. Вкладка ***Advanced*** містить такі параметри (рис. 9.4):

- ***Input file (Вхідний файл)*** – вибір варіанта вхідних даних ***Raw Data (Таблиця даних)*** або ***Correlation Matrix (Кореляційна матриця)***;
- ***Advanced options (stepwise or ridge regression) (Розширені опції (ступінчаста чи гребнева регресія)***;
- ***Review descriptive statistics, correlation matrix (Переглянути описові статистики, кореляційна матриця)***;
- ***Extended precision computation (Підвищена точність обчислень)***;
- ***Batch processing/reporting (Пакетна обробка даних/звіт)***.

Якщо вибрати ***Review descriptive statistics, correlation matrix*** і натиснути ***OK***, то з'явиться вікно ***Review Descriptive Statistics (Перегляд описових статистик)*** (рис. 9.5).

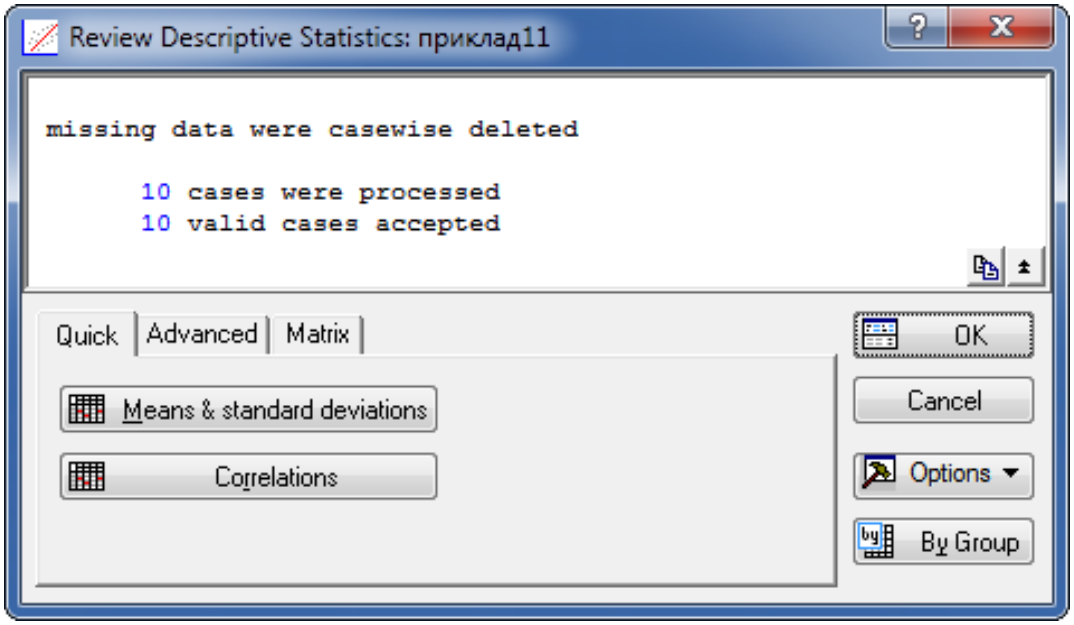


Рис. 9.5. Діалогове вікно ***Review Descriptive statistics***

На вкладці ***Quick*** (рис. 9.5) можна вибрати для аналізу ***Means & standard deviation (Середнє значення та середньоквадратичне відхилення)***, ***Correlations (Коефіцієнти кореляції)***. На вкладці ***Advanced*** можна ще вибрати ***Covariances (Коваріації),*** діаграми розмаху та матричні графіки розсіювання. Побудовані графіки дозволяють візуально перевірити розподіли на наявність "викидів", які можуть істотно вплинути на розташування кривої регресії.

2. Додаткові можливості здійснення кореляційно-регресійного аналізу в системі STATISTICA

У діалоговому вікні ***Multiple Regression Results*** на вкладці ***Advanced*** знаходяться кнопки аналізу, що дозволяють побудувати (рис. 9.6):

- ***Summary: Regression results –*** таблицю результатів регресійного аналізу (рис. 9.3);
- ***ANOVA (Overall goodness of fit)*** – таблицю дисперсійного аналізу (рис. 9.7);
- ***Covariance of coefficients (Коефіцієнти коваріації)*** – таблицю коефіцієнтів коваріації;
- ***Current sweep matrix*** – матрицю статистичного зв'язку між коефіцієнтами кореляції;
- ***Partial correlations*** – таблицю частинних коефіцієнтів кореляції (рис. 9.8);
- ***Redundancy*** – аналіз надлишковості;
- ***Stepwise regression summary*** – поетапний регресійний аналіз;
- ***ANOVA adjusted for mean*** – ANOVA з поправкою на середнє.

Таблиця частинних коефіцієнтів кореляції містить оцінки b*, частинні коефіцієнти кореляції, напівчастинні коефіцієнти кореляції (***Semipart Cor***.), толерантності (1–R-square) (***Tolerance***), коефіцієнти детермінації (***R-square***), значення *t*-критерію і рівні значущості *p* – імовірності відхилення гіпотези про значущість частинного коефіцієнта кореляції.

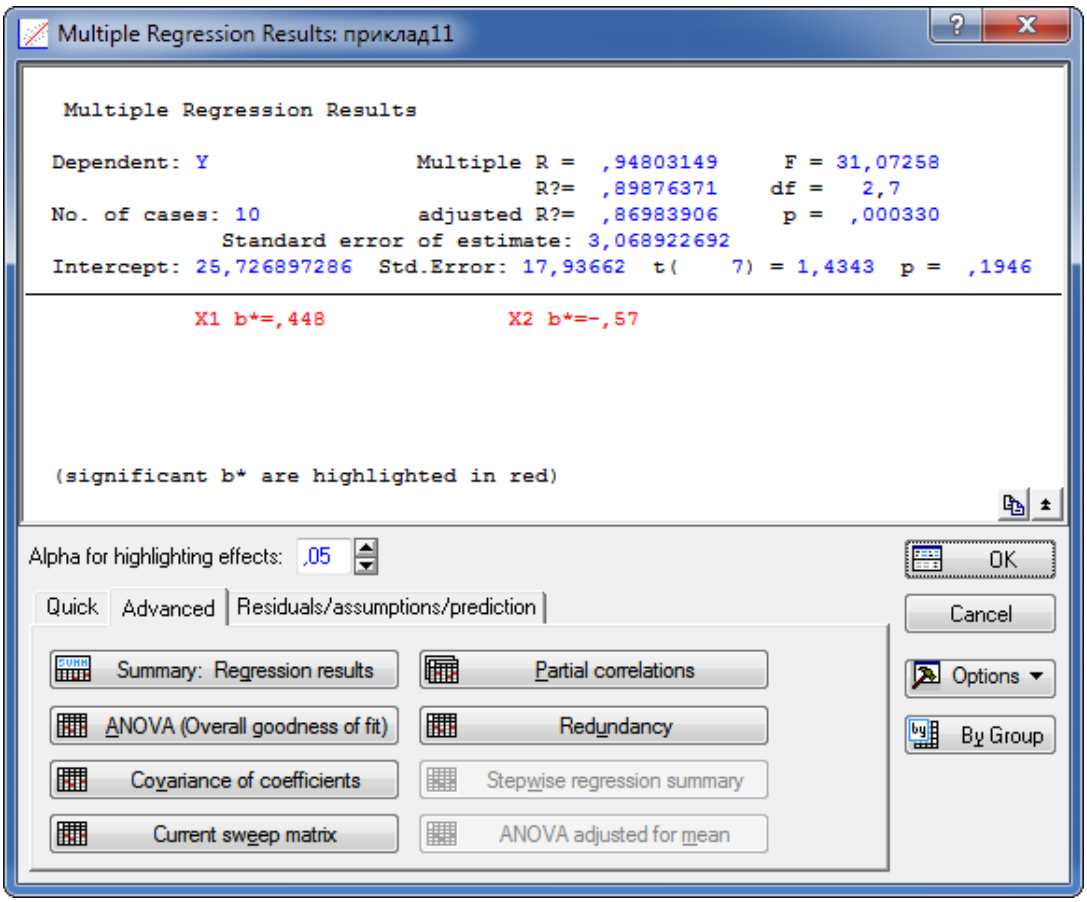


Рис. 9.6. Вкладка ***Advanced*** діалогового вікна ***Multiple Regression Results***

Effect	Analysis of Variance; DV: Y (Spreadsheet1_(Recovered))				
	Sums of Squares	df	Mean Squares	F	p-value
Regress.	585,3010	2	292,6505	31,07258	0,000330
Residual	65,9280	7	9,4183		
Total	651,2290				

Рис. 9.7. Таблиця аналізу ***ANOVA***

Variable	Variables currently in the Equation; DV: Y (Spreadsheet1)						
	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(7)	p-value
X1	0,447815	0,693495	0,306267	0,467738	0,532262	2,54672	0,038286
X2	-0,570488	-0,774978	-0,390165	0,467738	0,532262	-3,24436	0,014167

Рис. 9.8. Таблиця частинних коефіцієнтів кореляції

Важливим етапом регресійного аналізу є аналіз залишків (випадкових відхилень). Щоб провести аналіз залишків необхідно перейти на вкладку ***Residuals/assumptions/prediction (Залишки/припущення/прогноз)*** діалогового вікна ***Multiple Linear Regression***. Далі потрібно натиснути кнопку ***Perform residual analysis (Виконати аналіз залишків)***, у результаті чого відкриється робоче вікно ***Residuals Analysis (Аналіз залишків)*** (рис. 9.9).

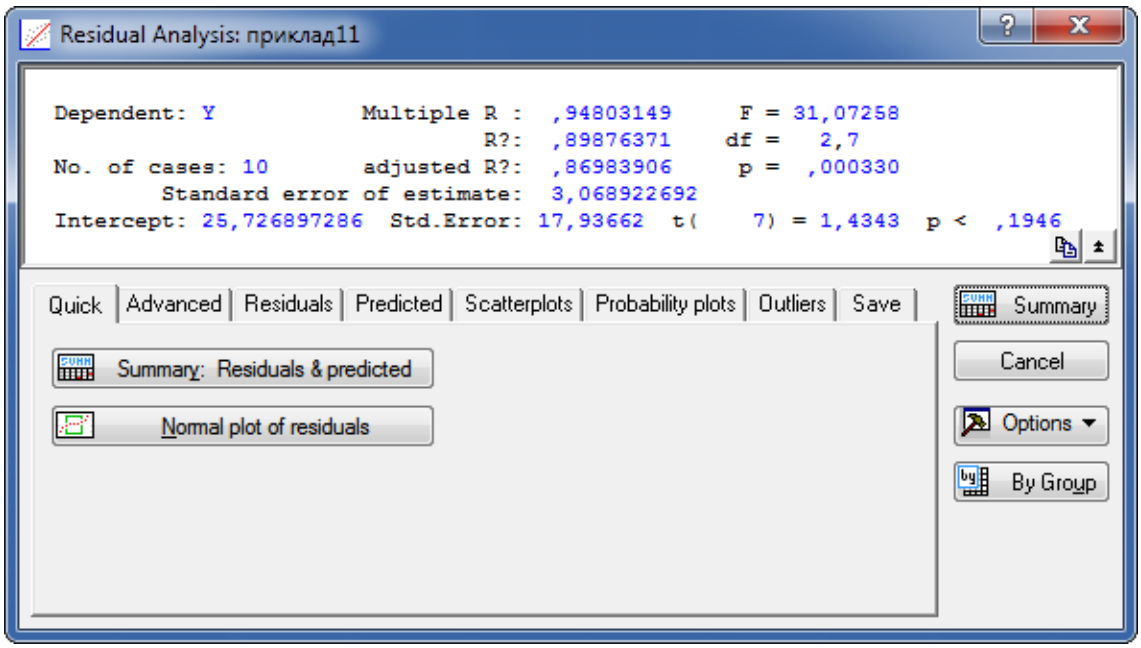
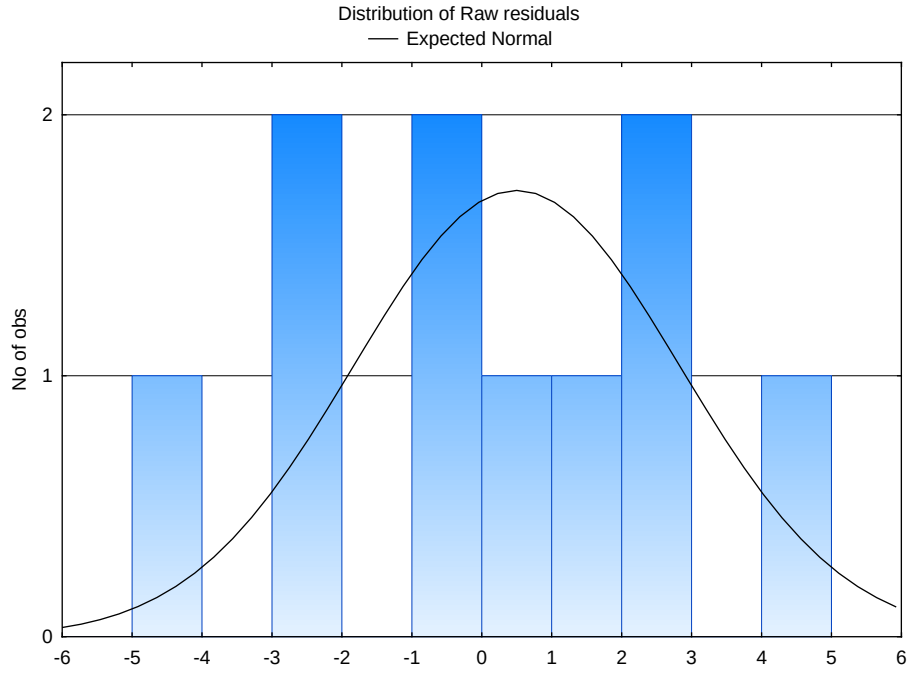


Рис. 9.9. Діалогове вікно ***Residuals Analysis***

При натисканні на кнопку ***Durbin-Watson statistic (Статистика Дарбіна-Уотсона)***, яка доступна на вкладці ***Advanced*** діалогового вікна (рис. 9.9), розраховується статистика, на основі якої можна перевірити наявність чи відсутність автокореляції між залишками для сусідніх спостережень.

У діалоговому вікні (рис. 9.9) можна перевірити залишки на нормальність розподілу, натискуючи на кнопку ***Histogram of residuals (Гістограма залишків)*** на вкладці ***Residuals*** або на кнопку ***Normal plot of residuals (Графік залишків і нормального розподілу)*** на вкладці ***Quick*** (рис. 9.10)**.** Якщо залишки розподілені нормально, то вони будуть знаходитися близько коло червоної, апроксимуючої лінії.



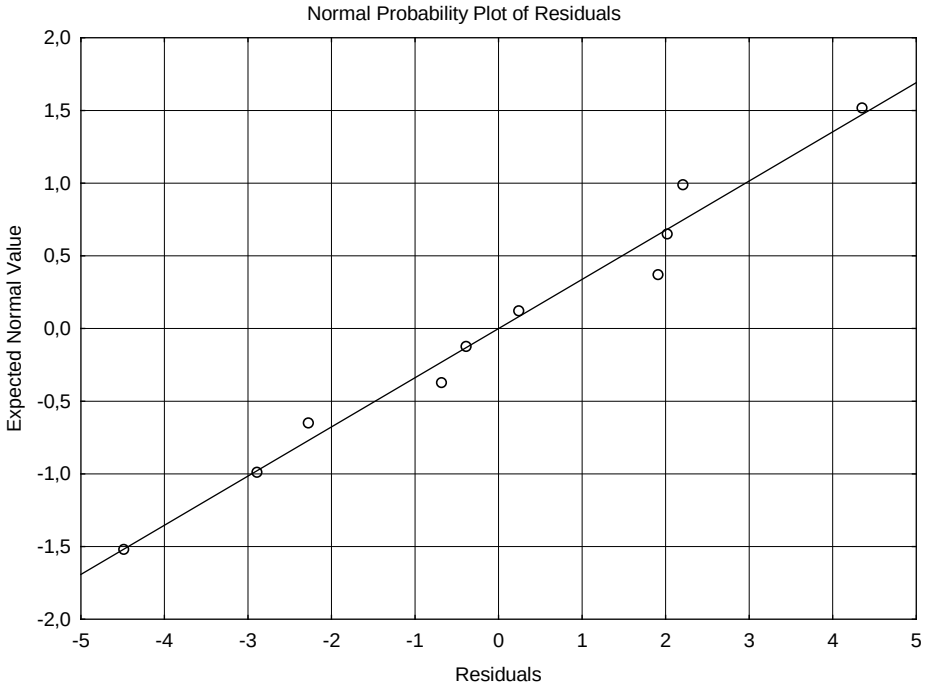


Рис. 9.10. Графіки розподілу залишків

Перевірка якості рівняння регресії здійснюється за допомогою гіпотез H_0 і H_1:

$$\begin{cases} H_0: \text{відсутній кореляційний зв'язок між досліджуваними змінними} \\ \quad \text{(побудоване рівняння регресії незначуще),} \\ H_1: \text{наявний кореляційний зв'язок між досліджуваними змінними} \\ \quad \text{(побудоване рівняння регресії значуще).} \end{cases}$$

Використовуючи алгоритм перевірки гіпотез, потрібно знайти спостережуване значення критерію F та критичне значення $F_{\theta;l_1;l_2}$, знайдене з таблиці F-розподілу. У системі **STATISTICA** спостережуване значення обчислюється, якщо вибрати функцію ***ANOVA*** у діалоговому вікні ***Multiple Regression Results*** (рис. 9.7), а також у діалоговому вікні ***Multiple Regression Results*** (рис. 9.2). Критичне значення $F_{\theta;l_1;l_2}$ можна знайти використовуючи ймовірннісний калькулятор ***Probability Distribution Calculator***, задавши відповідні параметри θ; $l_1 = k-1$; $l_2 = n-k$ (n – кількість спостережень, k – кількість оцінювальних параметрів регресії).

Якщо побудована регресійна модель адекватна (значуща), то можна визначити прогнозне значення результативної ознаки (залежної змінної) на основі побудованої регресійної моделі. Для цього потрібно у вікні ***Multiple Regression Results*** вибрати вкладку ***Residuals/assumptions/prediction*** і натиснути на кнопку ***Predict dependent variable (Спрогнозувати залежну змінну)***. Далі у відкритому вікні ***Specify values for indep.vars (Задати значення для незалежних змінних)*** (рис. 9.11) у полях змінних указати нові значення залежних змінних. Натиснувши ***OK*** у діалоговому вікні ***Specify values for indep.vars,*** отримаємо таблицю результатів прогнозування (рис. 9.12). У таблиці вказано прогнозоване (***predicted***) значення з 95%-м довірчим інтервалом.

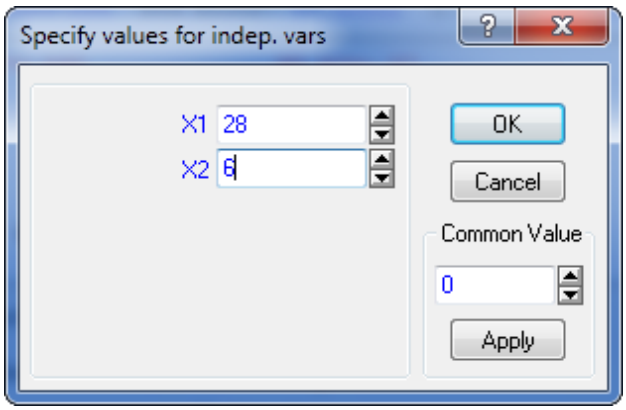


Рис. 9.11 Діалогове вікно ***Specify values for indep.vars***

Variable	Predicting Values for (приклад11 variable: Y b-Weight	Value	b-Weight * Value
X1	1,12819	28,00000	31,5893
X2	-3,16480	6,00000	-18,9888
Intercept			25,7269
Predicted			38,3274
-95,0%CL			32,0114
+95,0%CL			44,6434

Рис. 9.12. Таблиця результатів прогнозування

3. Типовий приклад

На основі вибіркових даних (табл. 9.1) побудувати вибіркове рівняння регресії, знайти основні кореляційні характеристики та здійснити їх аналіз, оцінити на 5%-му рівні значущість побудованої регресійної моделі.

Таблиця 9.1

Обсяг випуску, (млн. грн.)	*1,5*	*1,8*	*2,0*	*2,2*	*2,3*	*2,6*	*3,0*	*3,1*	*3,5*	*3,8*
Вартість основних фондів млн. грн.)	*3,9*	*4,4*	*3,8*	*3,5*	*4,8*	*4,3*	*7,0*	*6,5*	*6,1*	*8,2*

Розв'язування. *Скористаємося модулем* ***Multiple regression****. Вкажемо залежну та незалежну змінні (рис. 9.13).*

Щоб переглянути основні описові статистики та коефіцієнти кореляції треба у діалоговому вікні вказати ***Review descriptive statistics, correlation matrix.*** *У вікні, що з'явилося виберемо* ***Means & standard deviation*** *та* ***Correlations.*** *Результати подані на рис. 9.14. Далі перейдемо до* ***Multiple Regression Results*** *(рис. 9.15).*

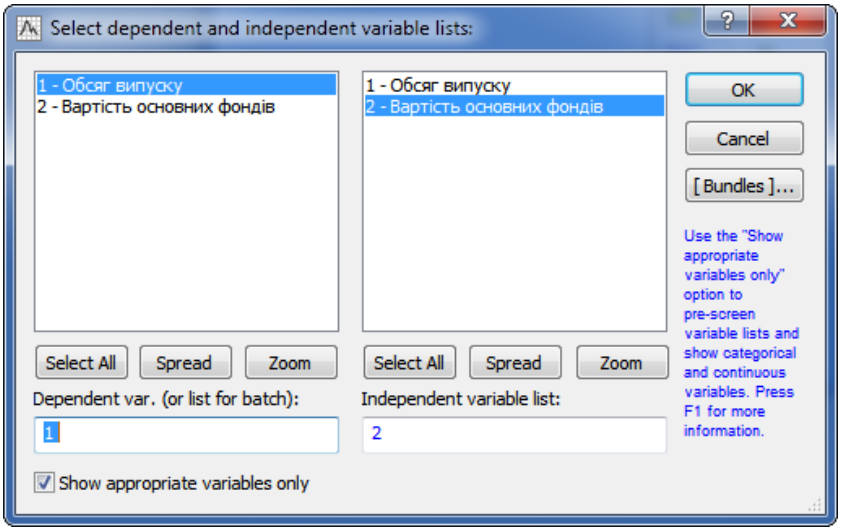


Рис. 9.13. Специфікація моделі

Variable	Means and Standard De Means	Std.Dev.	N
Вартість основних фондів	5,250000	1,593912	10
Обсяг випуску	2,580000	0,753953	10

Variable	Correlations (Spreadsheet17) Вартість основних фондів	Обсяг випуску
Вартість основних фондів	1,000000	0,878360
Обсяг випуску	0,878360	1,000000

Рис. 9.14. Результати описової статистики та кореляційна матриця

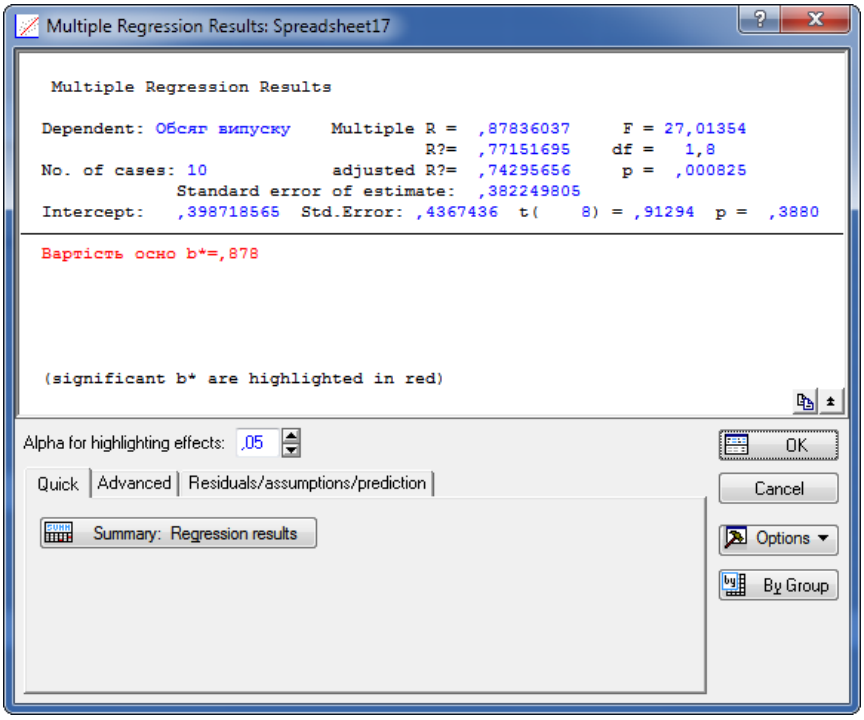


Рис. 9.15. Діалогове вікно результатів регресійного аналізу

Як видно з інформаційної частини, $R^2 \approx 0.77$, $F_{факт} \approx 27.01$ *та коефіцієнт регресії* b *значимий (підсвічений червоним). Для перевірки гіпотези про наявність лінійного зв'язку між обсягом випуску продукції та вартістю основних фондів розрахуємо за допомогою* ***Probability Distribution Calculator*** *критичне значення* $F_{табл} = 5{,}317655$ *(рис. 9.16) і порівняємо його з розрахованим.*

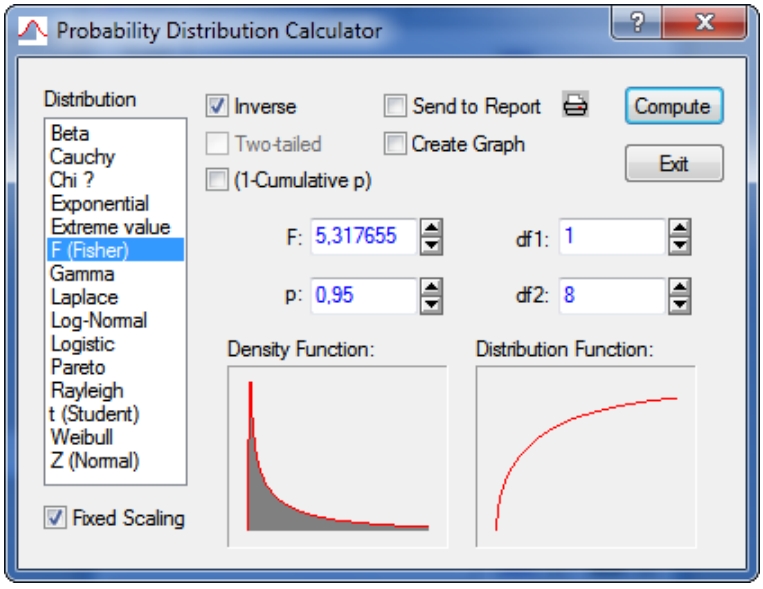


Рис. 9.16. Знаходження критичного значення F-критерію

Як бачимо, $F_{факт} > F_{табл}$, *отже між змінними існує лінійний зв'язок.*

Таблиця регресійного аналізу матиме вигляд, поданий на рис. 9.17.

N=10	Regression Summary for Dependent Variable: Обсяг випуску (R= ,87836037 R?= ,77151695 Adjusted R?= ,74295656 F(1,8)=27,014 p<,00082 Std.Error of estimate: ,38225					
	b*	Std.Err. of b*	b	Std.Err. of b	t(8)	p-value
Intercept			0,398719	0,436744	0,912935	0,387961
Вартість основних фондів	0,878360	0,168998	0,415482	0,079940	5,197455	0,000825

Рис. 9.17. Результати регресійного аналізу

Перевіримо залишки. Перейдемо до ***Residuals Analysis.*** *Побудуємо графік залишків (рис. 9.16). Для цього необхідно натиснути кнопку* ***Normal plot of residuals*** *на вкладці* ***Residuals.***

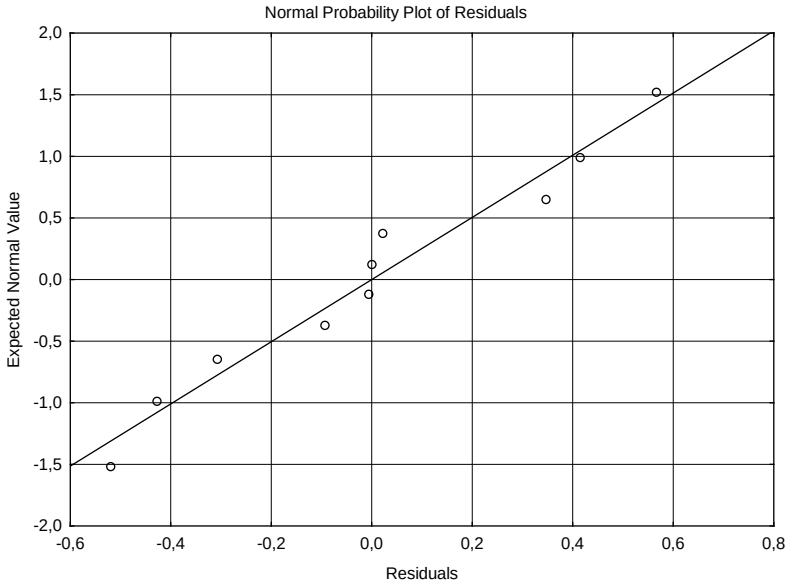


Рис. 9.16. Графік залишків

Очевидно, залишки розподілені нормально.

Отже, побудоване рівняння регресії є значуще і має вигляд $y = 0.3987 + 0.4155x$, *де* y – *усереднений обсяг випуску (млн. грн.),* x – *вартість основних фондів (млн. грн.).*

Завдання для самостійної роботи

9.1. У таблиці 9.2 наведено статистичні дані показників виробничо-господарської діяльності 50 промислових підприємств за деякий період часу.

Таблиця 9.2

№ з/п	Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	12	901	50	81	90	956	62	51	550	915	94	2
2	56	956	50	134	110	905	134	56	612	912	105	1
3	956	26	950	206	162	56	206	994	290	956	195	6
4	945	802	950	305	202	803	305	945	431	931	206	4
5	125	815	150	362	245	816	362	125	407	974	204	3
6	168	802	150	492	280	802	492	168	595	925	285	8
7	138	814	150	402	301	814	402	138	614	912	304	4
8	195	880	850	465	340	880	465	195	713	956	350	6
9	802	875	850	509	390	875	509	802	814	990	360	7
10	856	812	850	546	330	812	546	856	640	964	382	6
11	845	801	250	582	401	805	582	845	54	901	401	2
12	875	700	250	505	405	700	505	875	184	903	481	5
13	265	720	250	625	495	720	625	265	269	943	471	9
14	284	760	250	685	471	760	682	284	318	998	471	1
15	275	714	750	631	500	714	631	275	407	64	461	4
16	294	712	750	642	501	712	642	294	595	91	501	11
17	220	762	750	625	503	760	625	220	614	102	508	16
18	212	714	750	758	510	716	758	212	713	130	562	15
19	264	774	350	758	598	774	758	264	512	186	582	17
20	209	754	350	728	547	754	728	209	992	110	574	14
21	764	739	350	794	601	739	794	764	54	801	532	14
22	747	761	350	764	601	761	764	747	184	845	601	18
23	758	754	350	725	608	754	725	758	269	875	642	15
24	714	602	650	803	671	602	803	714	318	831	642	12
25	724	642	650	816	701	642	816	724	407	825	685	17
26	770	682	650	894	764	691	894	770	595	896	663	25
27	797	645	650	857	718	645	857	797	614	804	656	21
28	703	631	650	956	805	631	956	709	713	874	696	25
29	302	675	450	948	864	675	948	302	814	858	705	26
30	312	646	450	81	906	646	61	312	992	882	754	28
31	333	501	450	134	90	501	134	333	54	880	774	29
32	312	502	450	206	110	502	206	312	184	810	777	24
33	345	532	450	305	162	532	305	345	269	830	787	21
34	395	550	450	362	202	550	362	395	318	864	795	28
35	345	598	450	492	245	598	492	345	407	825	735	31
36	384	575	550	402	280	575	402	384	595	885	714	31
37	375	401	550	465	301	401	465	375	614	818	802	35
38	346	441	550	509	340	441	509	346	713	871	845	35
39	333	484	550	546	390	484	546	333	814	802	865	38
40	378	465	550	582	330	465	582	378	992	836	898	36
41	348	487	550	505	401	487	505	348	591	890	888	33
42	356	302	550	625	405	302	625	356	184	812	845	33
43	393	364	500	685	495	364	685	393	269	805	874	41
44	375	389	500	631	471	389	631	375	318	845	831	44
45	612	356	500	642	500	356	642	612	407	803	847	47

№ з/п	Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
46	645	165	500	625	501	165	625	645	595	845	801	46
47	636	198	500	758	503	198	758	636	614	802	905	45
48	604	256	500	758	510	256	758	604	713	864	914	45
49	679	230	500	728	598	230	728	679	814	812	975	47
50	659	140	500	794	547	140	794	659	992	885	969	52

Розглядаються такі показники:

Y_1 – продуктивність праці;

Y_2 – індекс зниження собівартості продукції;

Y_3 – рентабельність;

X_1 – трудомісткість одиниці продукції;

X_2 – питома вага робітників у складі ППП;

X_3 – питома вага купувальних виробів;

X_4 – коефіцієнт змінності устаткування;

X_5 – премії та винагороди на одного працівника;

X_6 – питома вага втрат від браку;

X_7 – фондовіддача;

X_8 – середньорічна чисельність ППП;

X_9 – середньорічна вартість ОФ;

1) Побудувати лінійну модель парної регресії відповідно до варіантів завдань (таблиця 9.3).

Обчислити основні кореляційні характеристики, знайти стандартні помилки регресії, оцінити на 5%-му рівні значущість побудованого рівняння регресії, визначити 95%-ві довірчі інтервали для прогнозованого індивідуального значення $y*$ залежної змінної при заданому значенні $x*$. Здійснити аналіз отриманих результатів.

2) Побудувати лінійну модель множинної регресії залежності результативної ознаки Y від чинників $X_1, ..., X_9$ відповідно до варіантів завдань.

Оцінити параметри лінійної множинної регресії. Обчислити кореляційну матрицю, множинні коефіцієнти детермінації R^2 та кореляції R, скоригований множинний коефіцієнт детермінації $\widehat{R}^2$, частинні коефіцієнти кореляції. Перевірити при рівні значущості $\theta = 0.05$ значущість коефіцієнтів регресії, значущість моделі в цілому, визначити 95%-ві довірчі інтервали для індивідуального значення залежної змінної при заданих значеннях пояснюючих змінних. Здійснити аналіз залишків.

Значення пояснюючих змінних для прогнозу результативної ознаки подані в табл. 9.4.

Таблиця 9.3

Варіанти завдань

№ варіанту	*Лінійна модель парної регресії*		*Лінійна модель множинної регресії*	
	Результативна ознака,	*Пояснююча змінна,*	*Результативна ознака,*	*Пояснюючі змінна,*
1	Y_3	X_7	Y_1	X_6, X_8, X_1, X_2
2	Y_2	X_1	Y_1	X_7, X_1, X_3
3	Y_2	X_8	Y_1	X_8, X_3, X_2
4	Y_2	X_9	Y_1	X_5, X_6, X_4
5	Y_3	X_5	Y_1	X_6, X_5, X_7, X_9
6	Y_2	X_6	Y_1	X_8, X_9, X_4
7	Y_2	X_4	Y_1	X_6, X_8, X_5

Таблиця 9.4

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
0,31	0,74	0,22	1,22	2,2	0,79	1,39	11795	78,11

9.2. За даними завдання 8.2 (лабораторна робота № 8) здійснити кореляційно-регресійний аналіз (визначити силу і напрям зв'язку між доходом і споживчими витратами, побудувати рівняння регресії й оцінити параметри, оцінити на 5%-му рівні значущість рівняння регресії, проаналізувати залишки, спрогнозувати значення результативної змінної при значенні доходу – 3500 грн.)

9.3. За даними завдання 8.3 (лабораторна робота № 8) здійснити кореляційно-регресійний аналіз (визначити силу і напрям зв'язку, побудувати вибіркове рівняння регресії, оцінити на 5%-му рівні значущість рівняння регресії, дослідити залишки, спрогнозувати значення залежної змінної при значенні вмісту фтору у воді – 4,5 мг/л.)

9.4. За даними завдання 8.4 (лабораторна робота № 8) здійснити кореляційно-регресійний аналіз (визначити силу і напрям зв'язку, побудувати вибіркове рівняння регресії, оцінити на 5%-му рівні значущість рівняння регресії, дослідити залишки, спрогнозувати значення залежної змінної при значенні етнічного різноманіття – 100).

9.5. За даними завдання 8.5 (лабораторна робота № 8) здійснити кореляційно-регресійний аналіз.

Лабораторна робота № 10
Критерій Стьюдента (*t*-критерій) для порівняння середніх значень двох вибірок. Непараметричні методи дослідження зв'язку між змінними

1. Основні теоретичні відомості про критерій Стьюдента

Одним із завдань у статистичних дослідженнях соціально-економічних явищ є порівняння середніх значень двох вибірок (наприклад, експериментальної та контрольної). Цю проблему можна вирішити за допомогою *t*-критерію Стьюдента (*t*-тест).

За допомогою даного тесту перевіряється нульова гіпотеза, яка полягає в тому, що обидві вибірки (групи) сформовані з однієї генеральної сукупності, тобто відмінності між середніми значеннями порівнюваних вибірок випадкові і не викликані дією досліджуваного фактора. Тест Стьюдента належить до групи параметричних методів аналізу. Його коректне застосування вимагає виконання трьох умов:

- обидві вибірки повинні бути незалежними, тобто властивості однієї з них ніяк не повинні бути пов'язані з властивостями іншої;
- обидві вибірки повинні мати нормальний закон розподілу або близький до нього;
- між дисперсіями вибірок не повинно бути статистично значущої різниці (однорідність дисперсій).

Перед застосуванням *t*-критерію необхідно знайти кількість ступенів вільності (*df*).

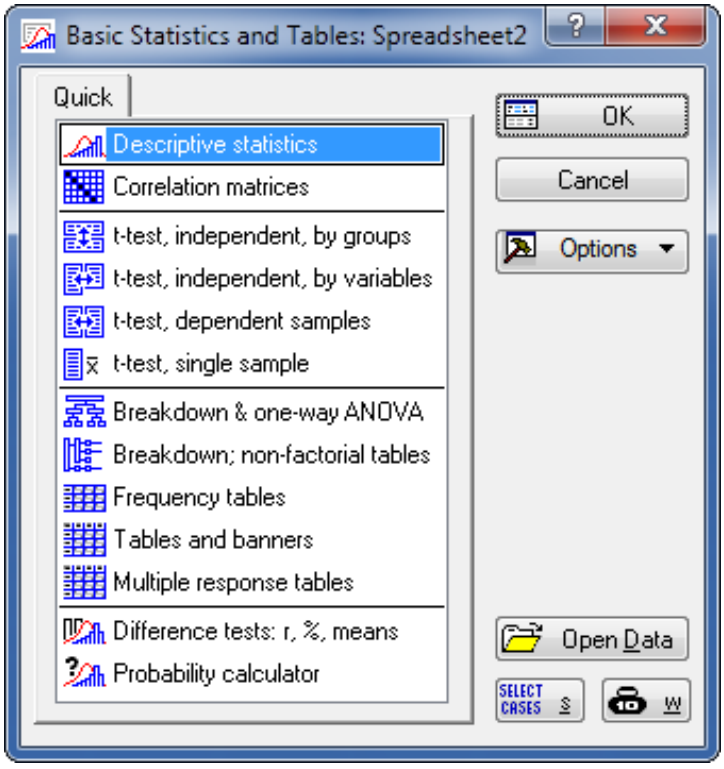


Рис. 10.1. Вікно модуля ***Basic Statistics/Tables***

Для реалізації критерію Стьюдента необхідно у вкладці ***Statistics*** групи ***Base*** вибрати або в меню ***Statistics*** модуль ***Basic Statistics/Tables***. Відкриється меню модуля (рис. 10.1), в якому *t*-критерій поданий чотирма процедурами (методами):

- ***t-test, independent, by variables (t-критерій для незалежних вибірок)*** – для порівняння середніх величин, отриманих за двома різними (незалежними) вибірками;
- ***t-test, independent, by groups (t-критерій для незалежних вибірок з групуючою змінною)*** – для порівняння середніх величин двох незалежних груп, отриманих з однієї

вибірки за допомогою групувальної змінної;

- ***t-test, dependent samples (t-критерій для залежних вибірок)*** – для порівняння середніх величин двох залежних груп (вибірок);
- ***t-test, single samples (прості вибірки)*** – для середніх, розрахованих по одній вибірці.

У перерахованих методах висувається нульова гіпотеза: середні значення двох вибірок (груп) рівні.

Для виконання *t*-тесту для незалежних вибірок необхідно у модулі ***Basic statistics/Tables*** вибрати ***t-test, independent, by variables*** (якщо дані не групуються) (рис. 10.2) або ***t-test, independent, by groups*** (якщо в таблиці даних є групувальна змінна) (рис. 10.3).

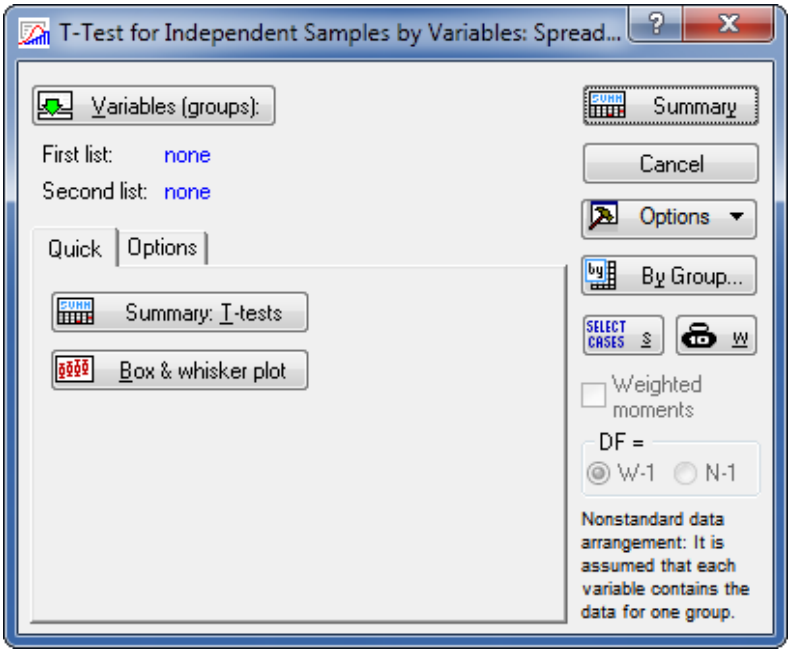


Рис. 10.2. Діалогове вікно t-***test, independent, by variables***

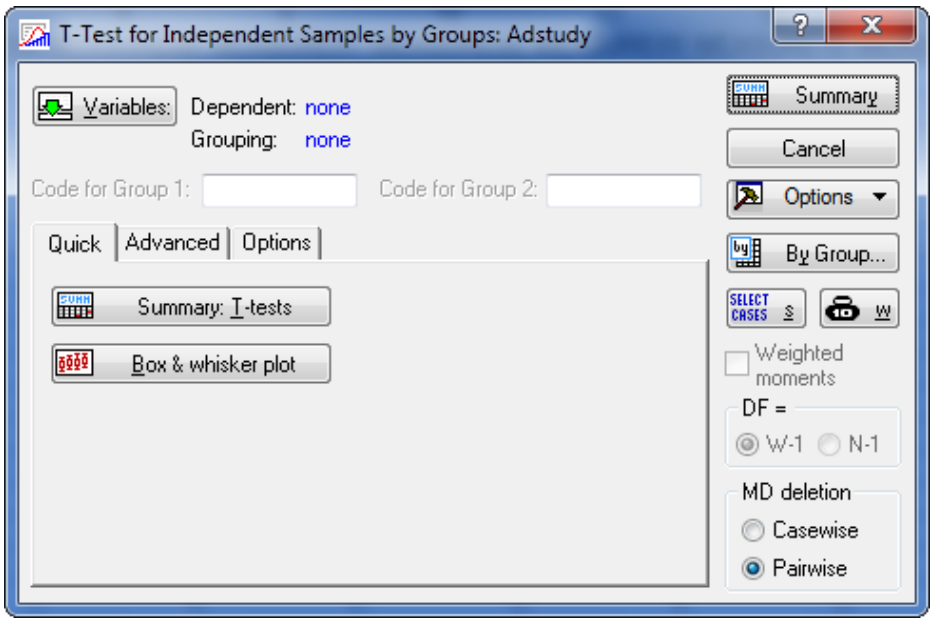


Рис. 10.3. Діалогове вікно ***t-test, independent, by groups***

Оскільки методи аналізу в обох *t*-тестах майже однакові, тому опишемо тест, де є групувальна змінна.

Для вибору змінних у вікні модуля ***t-test, independent, by groups*** необхідно натиснути

кнопку ***Variables*** (рис. 10.4) і вказати, яка змінна групувальна, а яка – залежна. Система автоматично вибере ознаки, за якими буде здійснено формування груп. Якщо необхідно вказати інші ознаки, можна двічі клацнути на полі ***Code for Group (Код для групи)*** і вибрати необхідну ознаку зі списку.

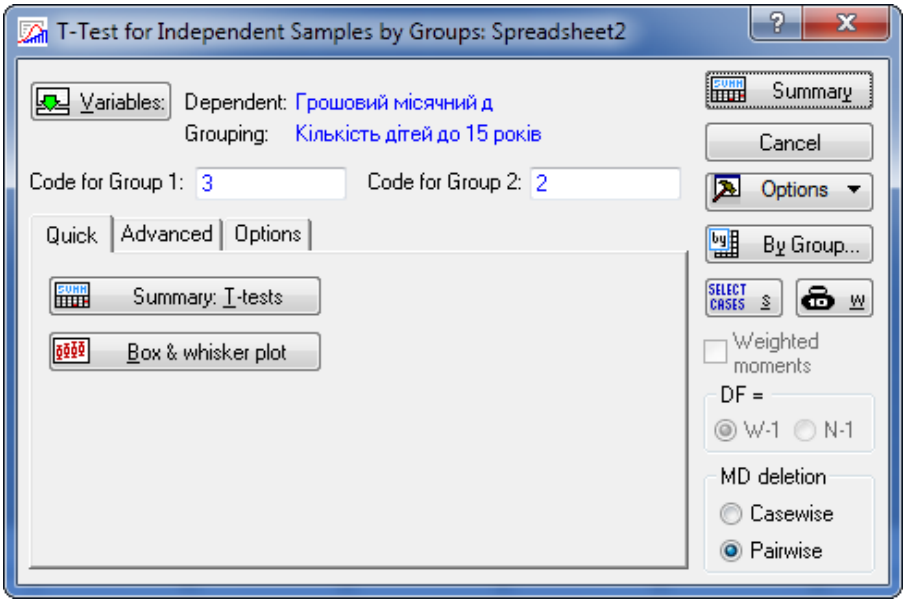


Рис. 10.4. Вибір змінних для перевірки *t*-тесту

Далі потрібно натиснути на кнопку ***Summary: T-tests (Результат: t-тест)***. У результаті система створить таблицю з результатами *t*-тесту (рис. 10.5).

T-tests; Grouping: Кількість дітей до 15 років (Spreadsheet2)
Group 1: 3
Group 2: 2

Variable	Mean 3	Mean 2	t-value	df	p	Valid N 3	Valid N 2	Std.Dev. 3	Std.Dev. 2	F-ratio Variances	p Variances
Грошовий місячний д	1571,250	2101,400	-2,38226	7	0,048720	4	5	161,1094	416,0863	6,669985	0,150934

Рис. 10.5. Результат *t*-тесту для незалежних вибірок з групувальною змінною

Опишемо таблицю результатів (рис. 10.5):

- ***Mean (3)*** – середнє значення доходу для групи, групувальною ознакою якої є 3;
- ***Mean (2)*** – середнє значення доходу для групи, групувальною ознакою якої є 2;
- ***t-value*** – значення розрахованого t-критерію Стьюдента;
- ***df*** – число ступенів вільності;
- ***p*** – імовірність помилково відкинути нульову гіпотезу про відсутність відмінностей між середніми. Фактично, це найголовніший результат аналізу. Якщо $p \geq 0.05$, H_0 приймається, якщо $p < 0.05$, H_0 відкидається.
- ***Valid N (3)*** – обсяг вибірки з групувальною ознакою 3;
- ***Valid N (2)*** – обсяг вибірки з групувальною ознакою 2;
- ***Std. dev. (3)*** – середньоквадратичне відхилення вибірки з групувальною ознакою 3;
- ***Std. dev. (2)*** – середньоквадратичне відхилення вибірки з групувальною ознакою 2;
- ***F-ratio Variances*** – значення F-критерію Фішера, за допомогою якого перевіряється гіпотеза про рівність дисперсій у порівнюваних вибірках;
- ***p Variances*** – імовірність помилки для F-тесту Фішера (гіпотеза про рівність дисперсій приймається, якщо ***p Variances*** більше, ніж 0,05).

Якщо виникають сумніви щодо однорідності дисперсій можна використати додаткові тести, що вказані на вкладці ***Options***: ***Levene's test (Тест Левене)*** або ***Brown&Forsythe test***

(тест Брауна-Форсайта) (рис. 10.4).

Графічний аналіз рівності середніх можна провести за допомогою діаграм розмаху.

Із залежними вибірками дослідник має справу кожен раз, коли аналіз значень досліджуваної ознаки виконуються на одних і тих же об'єктах. t-тест для двох залежних вибірок використовують, щоб перевірити, чи відрізняються два стовпці значень ознаки з погляду середнього значення при умові, що значення у двох стовпцях утворюють пари. Така ситуація виникає, наприклад, в дослідженнях "до/після", де розглядається результат вимірювання деякої ознаки (оцінки в результаті тестування чи рейтингу) для кожного об'єкта як до, так і після деякого втручання (перегляд реклами, проведення лікування тощо). Для порівняння середніх у таких вибірках необхідно використати ***t-test, dependent samples*** (рис. 10.6).

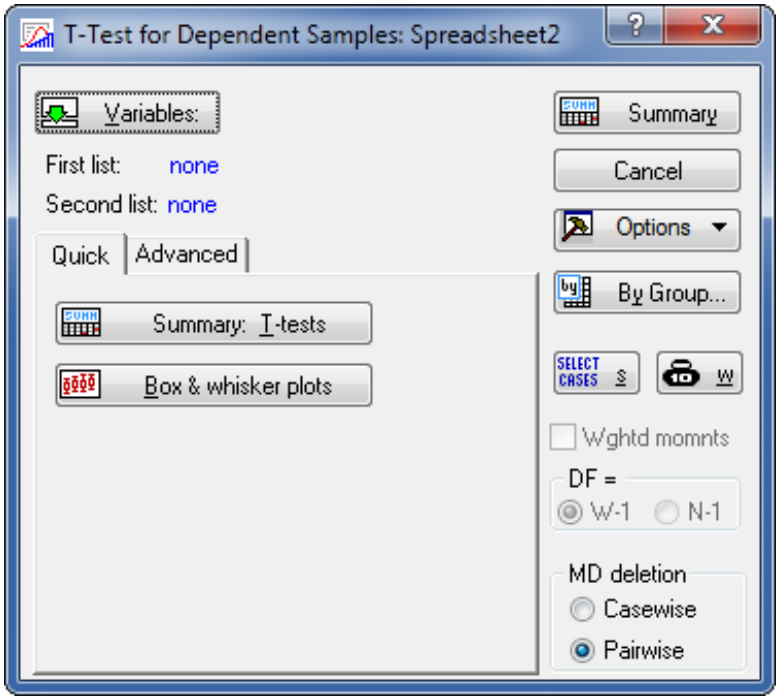


Рис. 10.6. Діалогове вікно модуля ***t-test, dependent samples***

Щоб вибрати змінні необхідно у діалоговому вікні натиснути кнопку ***Variables*** і вказати змінні, що беруть участь в аналізі. Далі потрібно натиснути кнопку ***Summary: T-tests***. У результаті з'явиться таблиця з результатами (рис. 10.7).

Variable	T-test for Dependent Samples (Spreadsheet2) Marked differences are significant at p < ,05000 Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p	Confidence -95,000%	Confidence +95,000%
Грошовий місячний д	1779,067	437,9352								
Середньодушові грошові витрати у місяць, грн.	387,933	128,7586	15	1391,133	379,1053	14,21198	14	0,000000	1181,192	1601,075

Рис. 10.7. Результати *t*-тесту для залежних вибірок

Таблиця результатів t-тесту для залежних вибірок містить:

- ***Mean*** – середні значення для кожної з порівнюваних груп;
- ***Std. dv.*** – середньоквадратичні відхилення для кожної групи;
- ***N*** – число спостережень;
- ***Diff.*** – різниця між середніми;
- ***Std. dv. diff.*** – середньоквадратичне відхилення для різниці між середніми;
- ***t*** – значення *t*-критерію;
- ***df*** – число ступенів вільності;
- ***p*** – імовірність помилково відкинути нульову гіпотезу про те, що середні величини

в порівнюваних групах не відрізняються. При наявності відмінностей, результати аналізу в **STATISTICA** зазвичай виділяються червоним кольором;

- 95%-м довірчі інтервали.

У деяких випадках виникає необхідність порівняти вибіркову середню не з іншою вибірковою середньою, а з певною константою. Для цього необхідно скористатися аналізом ***t-test for single means (t-тест для середніх, розрахованих по одній вибірці)***. Для цього аналізу треба вибрати ***t-test, single sample*** у модулі ***Basic Statistics/Tables***. У результаті з'явиться діалогове вікно (рис. 10.8).

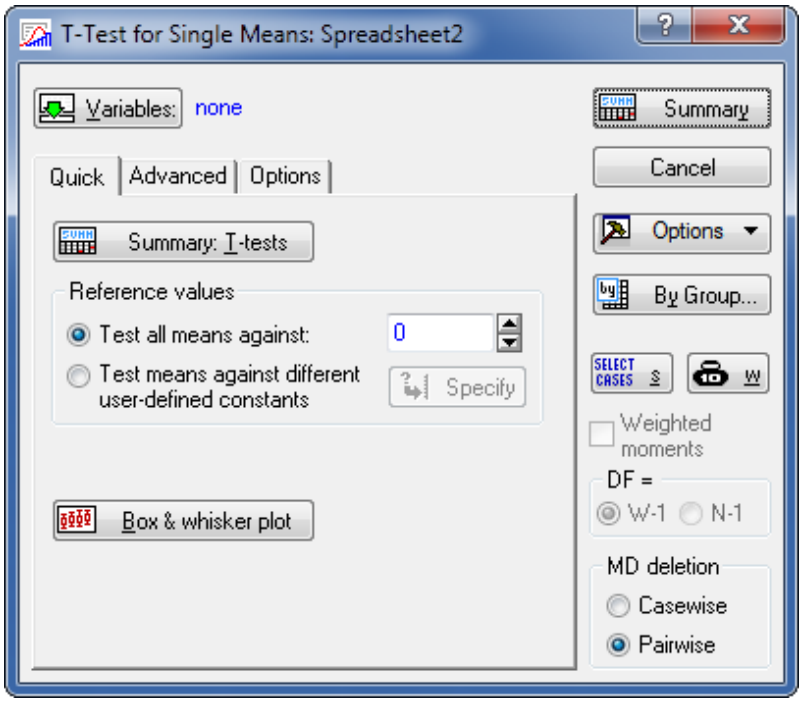


Рис. 10.8. Діалогове вікно ***t-test for single means***

Для вибору змінної (або змінних) необхідно натиснути кнопку ***Variables***. У такому випадку система порівняє всі вибірки з одним "контрольним" значенням. Останнє задається в полі ***Reference values (Контрольні значення)*** (рис. 10.8) і його треба внести навпроти опції ***Test all means against (Порівняти всі середні з)***. При необхідності внесені в аналіз змінні можна порівняти з кількома контрольними значеннями (це досягається шляхом активації опції ***Test means against user-defined constants (Порівняти середні з константами, заданими користувачем)***. Після натискання кнопки ***Summary: T-tests*** з'явиться таблиця з результатами аналізу (рис. 10.9).

	Test of means against reference constant (value) (Spreadsheet2)							
Variable	**Mean**	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
Грошовий місячний д	1779,067	437,9352	15	113,0744	1000,000	6,889861	14	0,000007

Рис. 10.9. Результати t-тесту для середніх, розрахованих по одній вибірці

У цій таблиці є така інформація:

- ***Mean*** – середнє значення;
- ***Std dv.*** – середньоквадратичне відхилення;
- ***N*** – обсяг вибірки;
- ***Std. err.*** – стандартна похибка;
- ***Reference constant*** – контрольне значення;

- ***t-value*** – значення розрахованого t-критерію Стьюдента;
- ***df*** – число ступенів вільності;
- ***p*** – ймовірність помилково відкинути нульову гіпотезу про те, що вибіркова середня не відрізняється від контрольної величини.

3. Непараметричні методи дослідження зв'язку між змінними

Непараметричні методи застосовуються для дослідження взаємозв'язку між якісними даними, поданих у номінальній шкалі або в порядковій шкалі (тобто у вигляді рангів), а також для кількісних даних у тому випадку, коли форма розподілу змінної невідома.

У системі **STATISTICA** є велика кількість методів аналізу та порівняння таких вибірок, а також для дослідження зв'язку між ними. Основні підходи до вибору непараметричного методу подано в таблиці 10.1.

Таблиця 10.1

Непараметричні методи аналізу вибірок

№ з/п	Початкова інформація	Зміст нульової гіпотези	Непараметричні методи
1	Дві незалежні вибірки обсягів n_1 і n_2.	Вибірки належать однорідним генеральним сукупностям.	1) критерій серій Вальда-Вольфовіца; 2) критерій Манна-Уїтні; 3) двовибірковий критерій Колмогорова-Смірнова.
2	Пари спостережень (x_i, y_i), $i = 1, 2, ..., n$ двох ознак X і Y, виміряних в порядкових або кількісних шкалах.	Ознаки X і Y некорельовані.	1) ранговий коефіцієнт кореляції Спірмена; 2) коефіцієнт кореляції Кендела.
3	k незалежних вибірок обсягів n_1, n_2, ..., n_k.	Вибірки належать однорідним генеральним сукупностям.	1) однофакторний дисперсійний аналіз Крускала-Уолліса; 2) медіанний критерій.
4	Дві пов'язані вибірки обсягів n.	Вибірки належать однорідним генеральним сукупностям.	1) критерій знаків; 2) критерій Вілкоксона.
5	k пов'язаних вибірок обсягів n.	Вибірки належать однорідним генеральним сукупностям.	1) двофакторний аналіз Фрідмана; 2) міри зв'язку – коефіцієнт конкордації Кендела.
6	Дві пов'язані вибірки обсягів n змінних X і Y, кожна з яких приймає два значення (0, 1; +, – тощо)	Ефект впливу відсутній.	критерій Макнімара.
7	k пов'язаних вибірок обсягів n змінних X_1, X_2, ..., X_k, кожна з яких приймає два значення.	Ефект впливу відсутній.	критерій Кокрена.
8	Вибірки двох випадкових змінних X і Y, кожна з яких приймає два значення.	X і Y незалежні.	аналіз таблиці спряженості 2×2 (точний критерій Фішера, критерій χ^2).

№ з/п	Початкова інформація	Зміст нульової гіпотези	Непараметричні методи
9	Вибірки двох змінних X і Y, що подані в номінальних шкалах. X приймає k значень, Y – r значень.	X і Y – незалежні.	аналіз таблиці спряженості $k \times r$ (критерій χ^2).

У системі **STATISTICA** є модуль ***Nonpametric Statistics***, який дозволяє реалізовувати описані методи. Для запуску модуля необхідно у вкладці ***Statistics*** у групі ***Base*** або в меню ***Statistics*** вибрати команду ***Nonpametrics (Непараметричні методи)***. Відкриється стартова панель модуля ***Nonpametric Statistics (Непараметричні статистики)*** (рис. 10.10).

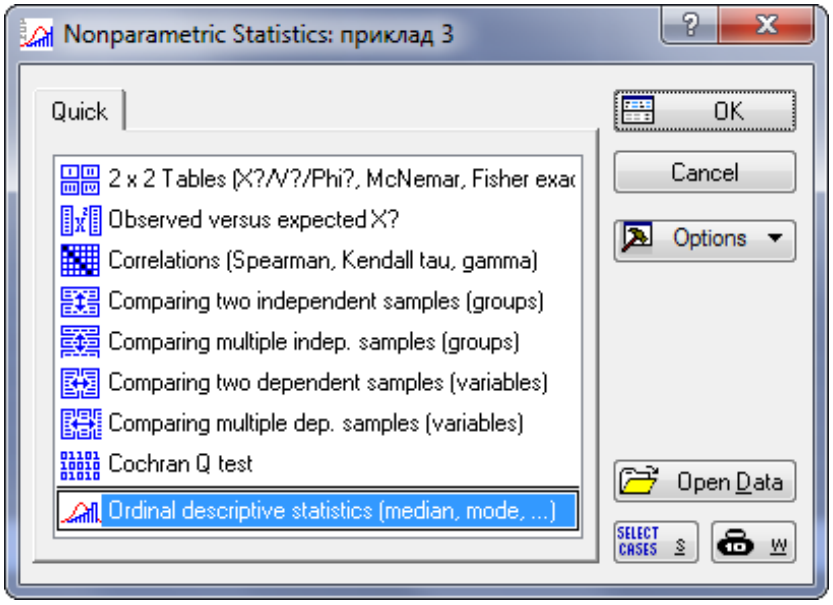


Рис. 10.10. Діалогове вікно ***Nonparametric Statistics***

Послідовно опишемо методи непараметричної статистики, що містяться в модулі ***Nonparametric Statistics***.

У процедурі ***2x2 Tables X?/V?/Ph?, McNemar, Fisher exact (Таблиці спряженості*** 2×2***, статистики*** χ^2***,*** φ***, критерій Макнімара, точний критерій Фішера)*** записуються частоти для двох ознак X і Y, кожна з яких набуває два значення: 0 і 1, “так” і “ні” тощо. Наприклад, якщо необхідно оцінити ставлення до певної програми чоловіків і жінок: 30 чоловікам подобається програма, 10 – ні, в той же час 20 жінкам програма подобається, 15 – ні, то таблиця спряженості сформується, як подано на рис. 10.11.

Опція 2×2 може бути використана як альтернатива кореляціям, якщо обидві змінні категоріальні.

Після натискання кнопки ***Summary***: ***2x2Table (Результати: Таблиці 2x2)*** з’явиться таблиця результатів (рис. 10.12).

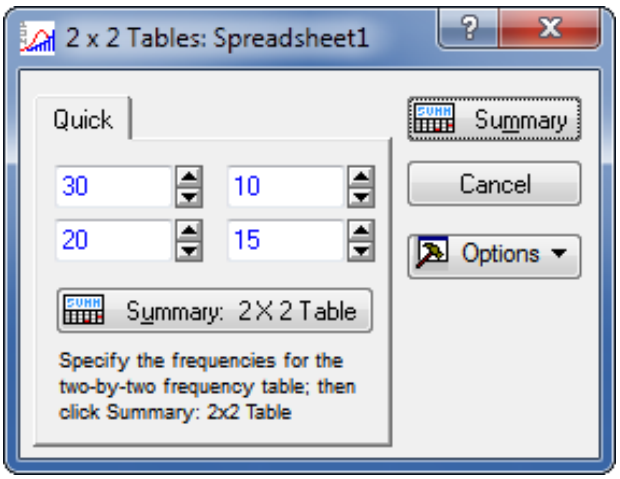


Рис. 10.11. Діалогове вікно таблиці спряженості 2×2

Додатково до стандартного критерію χ^2 Пірсона без поправки Йєтса і з нею та скорегованого χ^2 ***(V-square (V-квадрат))*** обчислюються такі статистики (рис. 10.12): ***Phi-square (φ^2)*** (міра зв'язку між номінальними або категоріальними змінними, значення яких не можна упорядкувати), значення якого змінюється від 0 (між змінними немає залежності) до 1 (між змінними є абсолютна залежність, тобто всі частоти розташовані на одній з діагоналей таблиці 2x2); ***точний критерій Фішера (Fisher exact)*** (обчислюється якщо сумарний обсяг вибірки невеликий ($n < 30$)), для якого розраховуються ***односторонні (one-tailed)*** і ***двосторонні (two-tailed)*** рівні значущості; ***McNemar Chi-square (Критерій значущості змін Макнімара)*** застосовується, якщо результативні дані – це дві пов'язані вибірки.

	2 x 2 Table (Spreadsheet1)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	30	10	40
Percent of total	40,000%	13,333%	53,333%
Frequencies, row 2	20	15	35
Percent of total	26,667%	20,000%	46,667%
Column totals	50	25	75
Percent of total	66,667%	33,333%	
Chi-square (df=1)	2,68	p= ,1017	
V-square (df=1)	2,64	p= ,1040	
Yates corrected Chi-square	1,94	p= ,1642	
Phi-square	,03571		
Fisher exact p, one-tailed		p= ,0820	
two-tailed		p= ,1413	
McNemar Chi-square (A/D)	4,36	p= ,0369	
Chi-square (B/C)	2,70	p= ,1004	

Рис. 10.12. Таблиця результатів методу ***2x2Table***

Процедура ***Observed versus expected X? (Статистика χ^2 для порівняння спостережуваних і очікуваних частот)*** діалогового вікна ***Nonparametric Statistics*** використовує статистику χ^2 для перевірки узгодженості спостережуваних і очікуваних частот. Процедура пропонує користувачу ввести дві змінні: одна містить очікувані, інша – спостережувані частоти.

Після натискання кнопки ***Summary*** з'явиться таблиця з результатами (рис. 10.13).

Case	Observed vs. Expected Frequencies (Spreadsheet3) Chi-Square = 491,2350 df = 12 p = 0,000000 NOTE: Unequal sums of obs. & exp. frequencies **observed Врожайність 1**	expected Врожайність	O - E	(O-E)**2 /E
C: 1	2090,00	1920,00	170,000	15,0521
C: 2	2252,00	2020,00	232,000	26,6455
C: 3	2360,00	2060,00	300,000	43,6893
C: 4	2320,00	1960,00	360,000	66,1224
C: 5	2240,00	1960,00	280,000	40,0000
C: 6	2100,00	2140,00	-40,000	0,7477
C: 7	2296,00	1980,00	316,000	50,4323
C: 8	2249,00	1940,00	309,000	49,2170
C: 9	2321,00	1790,00	531,000	157,5201
C: 10	2368,00	2250,00	118,000	6,1884
C: 11	2205,00	2410,00	-205,000	17,4378
C: 12	2261,00	2260,00	1,000	0,0004
C: 13	2400,00	2200,00	200,000	18,1818
Sum	29462,00	26890,00	2572,000	491,2350

Рис. 10.13. Результати порівняння спостережуваних і очікуваних частот

Зазначимо, що в нижній частині таблиці результатів показано загальне число випадків ***Sum (Сума)***; різниці між спостережуваними й очікуваними значеннями подані в третьому стовпчику, квадрати різниць, поділені на очікувані значення (параметри χ^2), – в четвертому стовпці.

Процедура ***Correlations (Spearman, Kendall tau, gamma)*** діалогового вікна ***Nonparametric Statistics*** дозволяє обчислити три різні альтернативи коефіцієнту кореляції Пірсона: кореляцію Спірмена, Кендела і гамма. Після вибору методу на екрані з'явиться діалогове вікно, в якому можна вибрати змінні і тип відображення коефіцієнтів кореляції: ***Detailed report (Детальний звіт)***, ***Square matrix (Квадратна матриця)*** і ***Matrix of two lists (Матриця, яка формується з двох списків)***, що вибирається у списку ***Compute (Обчислити)*** (рис. 10.14).

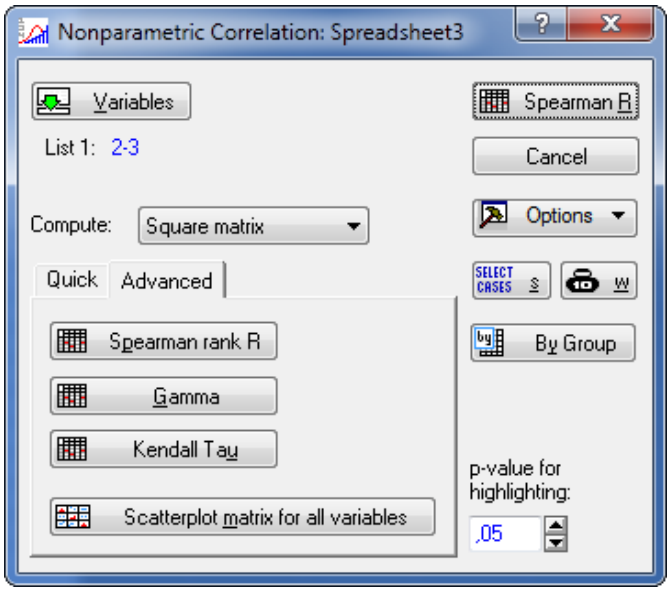


Рис. 10.14. Діалогове вікно ***Correlations (Spearman, Kendall tau, gamma)***

На вкладці ***Advanced*** діалогового вікна рис. 10.14 необхідно вибрати потрібний коефіцієнт кореляції та натиснути відповідну кнопку. З'явиться таблиця з результатами аналізу, яка містить, залежно від вибраного коефіцієнта, ***Valid N (Число спостережень)***, ***Spearman R (Коефіцієнт кореляції Спірмена)***, ***t (N-2) (Значення критерію Стьюдента для***

числа ступенів вільності n-2), ***p–value –*** ймовірність помилки для нульової гіпотези про відсутність зв'язку між ознаками, ***Gamma (Коефіцієнт кореляції Гамма), Kendall tau (Коефіцієнт кореляції Кендела)*** та ***Z (Z статистика).***

Досліджуваний зв'язок можна візуалізувати за допомогою матричних графіків розсіювання (***Scatterplot matrix for all variables*** на рис. 10.14).

Процедура ***Comparing two independent samples (groups) (Порівняння двох незалежних вибірок (груп))*** діалогового вікна ***Nonparametric Statistics*** (рис. 10.15) дозволяє розрахувати критерій серій Вальда–Вольфовіца, критерій Манна-Уітні та двовибірковий критерій Колмогорова-Смірнова.

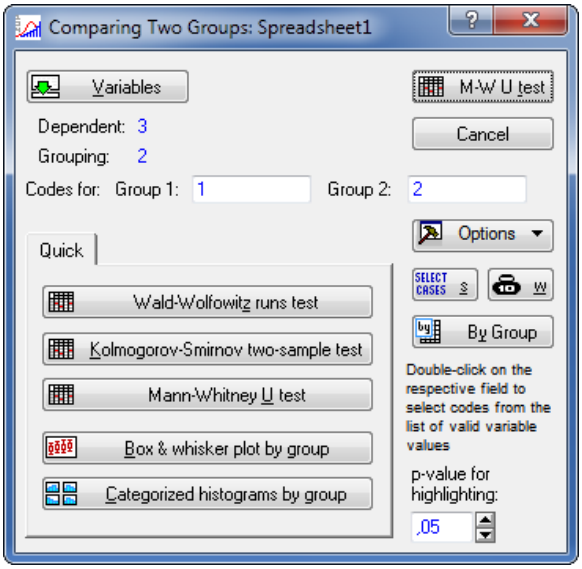


Рис. 10.15. Діалогове вікно ***Comparing two independent samples (groups)***

Wald-Wolfowitz runs test (Критерій серій Вальда-Вольфовіца) є непараметричною альтернативою t-критерію для незалежних вибірок. Файл даних повинен містити групувальну (незалежну) змінну, яка набуває принаймні два різних значення (коди), щоб однозначно визначити, до якої групи відноситься кожне спостереження у файлі даних. Змінні повинні бути виміряні в порядковій шкалі.

Mann-Whitney U-test (Критерій Манна-Уітні) аналогічний критерію серій Вальда-Вольфовіца, однак він порівнює не середні значення вибірок, а суми рангів по кожній з них.

А ***Kolmogorov-Smirnov two-sample test (двовибірковий критерій Колмогорова-Смірнова)*** порівнює емпіричні функції розподілу двох вибірок.

Указані вище критерії перевіряють гіпотезу про те, що дві незалежні вибірки сформовані з однієї і тієї ж генеральної сукупності.

У діалоговому вікні ***Nonparametric Statistics*** є процедура ***Comparing multiple indep. samples (groups) (Порівняння кількох незалежних вибірок (групи))*** (рис. 10.16), яка дозволяє провести ***Kruskal-Wallis ANOVA & Median test (ANOVA Крускала–Уолліса і медіанний тест)*** та ***Multiple comparisons of mean ranks for all groups (Множинне порівняння середніх рангів для всіх груп).*** Ці два тести є непараметричними альтернативами дисперсійного аналізу.

Критерій Крускала-Уолліса є узагальненням критерію Манна-Уітні на випадок, коли число вибірок $k > 2$ та перевіряє гіпотезу: k незалежних вибірок обсягів n_1, n_2, ..., n_k належать однорідним генеральним сукупностям. Цей критерій заснований на рангах, а не на середніх значеннях.

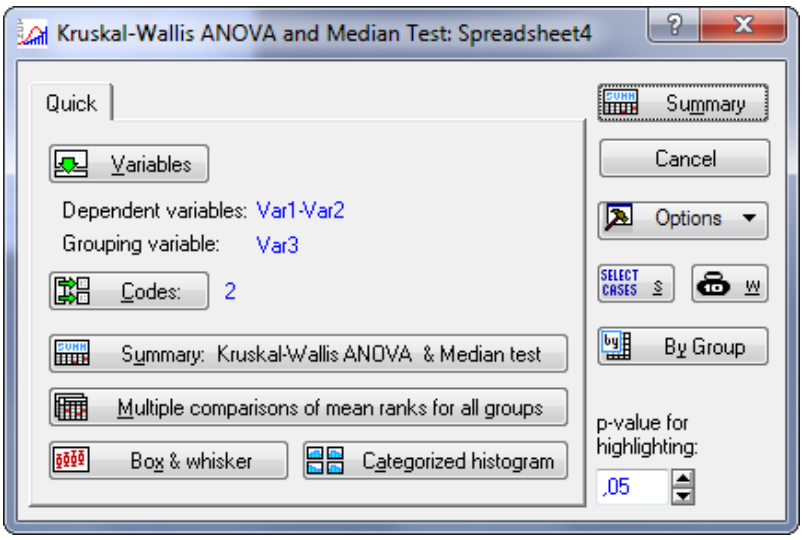


Рис. 10.16. Діалогове вікно ***Kruskal-Wallis ANOVA & Median test***

Медіанний критерій використовується для перевірки нульової гіпотези про те, що всі k вибірок, що отримані з генеральних сукупностей, мають рівні медіани.

Multiple comparisons of mean ranks for all groups використовується для обчислення апостеріорних порівнянь середніх рангів для всіх пар груп.

Процедура ***Comparing two dependent samples (variables) (Порівняння двох залежних вибірок (змінних))*** діалогового вікна ***Nonparametric Statistics*** (рис. 10.17) дозволяє розрахувати критерії знаків та Вілкоксона. При вибору даного методу можна обчислити:

- ***Sign Test (Критерій знаків)*** – це непараметрична альтернатива t-критерію для залежних вибірок і застосовується для перевірки гіпотези H_0 про однорідність генеральних сукупностей за попарнопозв'язаними вибіркам.
- ***Wilcoxon Matched Pairs Test (Критерій Вілкоксона)*** аналогічний критерію знаків, однак порівнюються не знаки, а ранги. Статистика Вілкоксона дорівнює найменшому значенню суми рангів різниць.

Процедура ***Comparing multiple dep. samples (variables) (Порівняння багатьох залежних вибірок (змінних))*** діалогового вікна ***Nonparametric Statistics*** дозволяє розрахувати ***Friedman ANOVA and Kendall's concordance (ANOVA Фрідмана та коефіцієнт конкордації, або згоди, Кендалла)***.

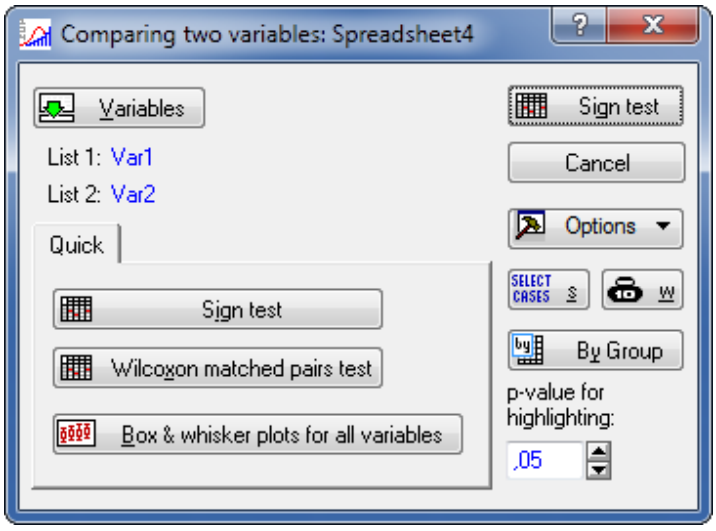


Рис. 10.17. Діалогове вікно ***Comparing two dependent samples (variables)***

ANOVA Фрідмана – це непараметрична альтернатива дисперсійному аналізу з повторними вимірюваннями. Коефіцієнт конкордації (згоди) Кендала – непараметричний коефіцієнт кореляції між двома змінними, коли число змінних більше двох.

Усі процедури порівняння графічно подаються за допомогою діаграм розмаху та, у випадку порівняння груп, категоризованих гістограм.

Cochran Q Test (Q-критерій Кохрена) діалогового вікна ***Nonparametric Statistics*** – це розширення критерію χ^2 Макнімара. Критерій перевіряє значущість відмінності декількох порівнюваних змінних, які приймають значення 0 або 1. Після вибору опції Q-критерій Кохрена у стартовій панелі ***Nonparametric Statistics*** програма запропонує визначити список змінних і коди, що ідентифікують (дві категорії або два фактора). Реалізація критерію в системі **STATISTICA** припускає, що змінні закодовані як одиниці і нулі, або коди, визначені користувачем, відповідно перетворяться в ці значення (тільки для даного аналізу, сам по собі файл не буде змінений).

При проведенні тестів щодо рівності середніх і дисперсій треба пам'ятати, якщо $p \geq 0,05$, H_0 приймається, якщо $p < 0,05$, H_0 відкидається.

4. Приклади типових завдань

1. *Двісті покупців магазину побутової техніки дали відповіді на питання: «Чи хочете ви купити кухонний комбайн нової марки?» до і після того як їм був показаний рекламний ролик. Частоти відповідей наведені в таблиці 10.2*

Таблиця 10.2

До	*Після*	
	Не бажаю придбати	*Бажаю придбати*
Бажаю придбати	*10*	*71*
Не бажаю придбати	*74*	*45*

Чи показують ці результати, що перегляд рекламного ролика ефективно вплинув на покупців?

Розв'язування. *Скористаємося модулем* ***2x2 Tables X?/V?/Ph?, McNemar, Fisher exact.*** *Таблиця спряженості матиме вигляд поданий на рис. 10.18. Результати розрахунків критерію – на рис. 10.19.*

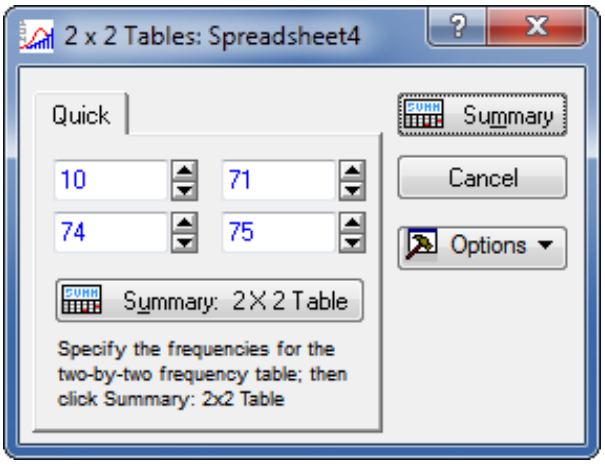


Рис. 10.18. Таблиця спряженості

	2 x 2 Table (Spreadsheet4)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	10	71	81
Percent of total	4,348%	30,870%	35,217%
Frequencies, row 2	74	75	149
Percent of total	32,174%	32,609%	64,783%
Column totals	84	146	230
Percent of total	36,522%	63,478%	
Chi-square (df=1)	31,52	p= ,0000	
V-square (df=1)	31,39	p= ,0000	
Yates corrected Chi-square	29,93	p= ,0000	
Phi-square	,13705		
Fisher exact p, one-tailed		p= ,0000	
two-tailed		p= ,0000	
McNemar Chi-square (A/D)	48,19	p= ,0000	
Chi-square (B/C)	,03	p= ,8681	

Рис. 10.19. Результати аналізу

*Оскільки вибірка велика, висновки можна зробити за χ^2 критерієм. Значення розрахованого χ^2 критерію більше за критичне значення 3,841459 (розраховане за допомогою **Probability Calculator**), p менше за 0,05, тому нульова гіпотеза відкидається, тобто можна вважати, що вибірки неоднорідні і реклама вплинула на покупців.*

***2.** Метод отримання випадкових чисел був застосований 250 разів, при цьому отримані такі результати (табл. 10.3)*

Таблиця 10.3

Число	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Частота появи	*27*	*18*	*23*	*31*	*21*	*23*	*28*	*25*	*22*	*32*

Чи можна вважати, що числа 0-9 з'являються з однією і тією ж імовірністю.

***Розв'язування.** Скористаємося модулем **Observed versus expected X?**. Якщо припущення вірне, то очікувана частота появи чисел дорівнює 25 (250/10). Запишемо число 25 як змінну 3 і порівняємо частоти. Результати подані на рис. 10.20.*

*Значення розрахованого χ^2 критерію менше за критичне значення 16,918978 (розраховане за допомогою **Probability Calculator**), p більше за 0,05, то нульова гіпотеза приймається, тобто можна вважати, що числа з'являються з однаковою частотою.*

	Observed vs. Expected Frequencies (Spreadsheet Chi-Square = 7,200000 df = 9 p = ,616305			
Case	**observed Частота появи**	expected Var3	O - E	(O-E)**2 /E
Var1	27,0000	25,0000	2,00000	0,160000
Var2	18,0000	25,0000	-7,00000	1,960000
Var3	23,0000	25,0000	-2,00000	0,160000
Var4	31,0000	25,0000	6,00000	1,440000
Var5	21,0000	25,0000	-4,00000	0,640000
Var6	23,0000	25,0000	-2,00000	0,160000
Var7	28,0000	25,0000	3,00000	0,360000
Var8	25,0000	25,0000	0,00000	0,000000
Var9	22,0000	25,0000	-3,00000	0,360000
Var10	32,0000	25,0000	7,00000	1,960000
Sum	250,0000	250,0000	0,00000	7,200000

Рис. 10.20. Результати порівняння

3. *Обсяги продаж у двох магазинів побутової техніки протягом 10 днів подані в таблиці 10.4 (в тис. грн.)*

Таблиця 10.4

X	*19*	*15*	*17*	*18*	*17*	*18*	*21*	*21*	*15*	*13*
У	*19*	*17*	*17*	*17*	*17*	*19*	*20*	*19*	*15*	*14*

Визначити величину зв'язку між обсягами продажів за допомогою коефіцієнтів рангової кореляції.

Розв'язування. *Скористаємося модулем* ***Correlations (Spearman, Kendall tau, gamma)*** *і обчислимо кореляцію Спірмена та Кендела. Результати подані на рис. 10.21.*

Spearman Rank Orde
MD pairwise deleted
Marked correlations a

Variable	X	У
X	1,000000	0,917138
У	0,917138	1,000000

а)

Kendall Tau Correlati
MD pairwise deleted
Marked correlations a

Variable	X	У
X	1,000000	0,858956
У	0,858956	1,000000

б)

Рис. 10.21. Коефіцієнти кореляції Спірмена (а) та Кендела (б)

З матриць кореляції випливає, що обсяги продажів мають тісний зв'язок.

4. *При вивченні іноземної мови у двох групах студентів використовувалися дві різні методики. Після вивчення частини курсу студенти обох груп написали диктант. Кількість помилок у диктанті подана в табл. 10.5.*

Таблиця 10.5

1 група	*31*	*26*	*33*	*11*	*13*	*5*	*18*	*1*	*2*	*16*	*17*	*23*	*20*	*21*	*9*			
2 група	*12*	*7*	*4*	*8*	*3*	*6*	*10*	*25*	*22*	*24*	*15*	*19*	*14*	*36*	*34*	*32*	*27*	*35*

Чи можна вважати, що застосування різних методик не приводить до суттєвих розбіжностей в результатах диктанту?

Розв'язування. *Порівняємо вибірки за допомогою* ***t-test, independent, by variables*** *(припустимо, що змінні мають нормальний розподіл). Вибір змінних зробимо як показано на рис. 10.22.*

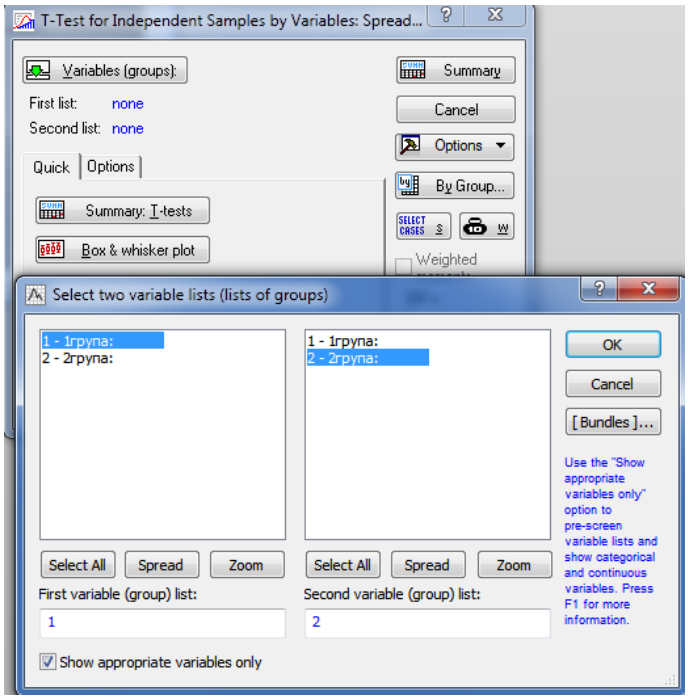


Рис. 10.22. Вибір змінних для t-критерію

З таблиці випливає, що розрахований t-test менший від критичного (1,695519) та незначущий (p > 0,05). Отже, за результатами тесту можна вважати: вибірки вибрані з однієї сукупності і різні методики не впливають на результати навчання.

Перевіримо вплив методик на навчання за допомогою непараметричної процедури ***Comparing two independent samples (groups).*** *Для цього виберемо як залежну змінну – кількість помилок, а групувальну – номер групи (рис. 10.24) і проведемо один із указаних у діалоговому вікні аналізів.*

Усі інші порівняння вибірок проводяться аналогічно.

Результат тесту поданий на рис. 10.23.

T-test for Independent Samples (Spreadsheet23)
Note: Variables were treated as independent samples

Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
1група: vs. 2група:	16,40000	18,50000	-0,566442	31	0,575172	15	18	9,752655	11,25768	1,332453	0,593842

Рис. 10.23. Результат порівняння вибірок

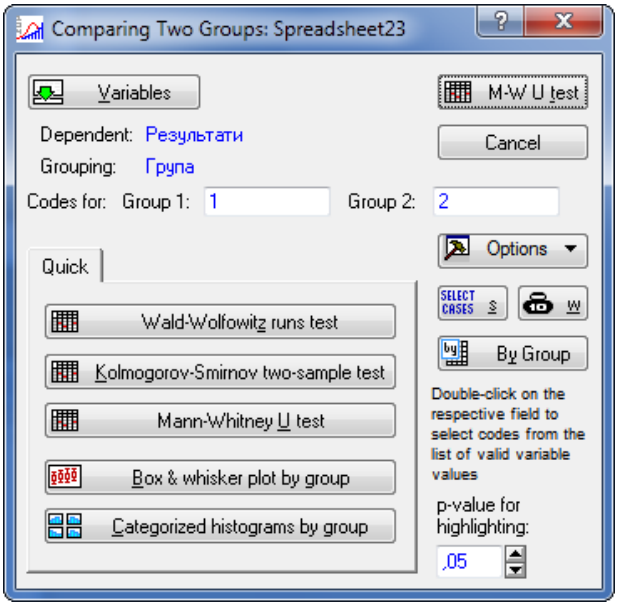


Рис. 10.24. Вибір змінних

Результати аналізу показують, що між вибірками немає різниці (критерій незначущий) (рис. 10.25).

Mann-Whitney U Test (Spreadsheet23)
By variable Група
Marked tests are significant at p <,05000

variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2	2*1sided exact p
Результати	240,0000	321,0000	120,0000	-0,524249	0,600106	-0,524249	0,600106	15	18	0,604760

*Рис. 10.25. Результати аналізу за допомогою непараметричної процедури (****Критерій Манна–Уітні****)*

Завдання для самостійної роботи

10.1. Досліджуються два виробничих процеси виготовлення поршневих кілець (табл. 10.6). Перевірити гіпотезу про однаковість відсотка браку в обох процесах.

Таблиця 10.6

Кільця	*Процес*	
	1	*2*
Придатні	195	149
Браковані	5	2

10.2. Під час епідемії грипу вивчалася ефективність щеплень проти цього захворювання. Отримані результати подані в таблиці 10.7. Визначити, чи ефективне щеплення.

Таблиця 10.7

Після щеплення		*Без щеплення*	
захворіли	не захворіли	захворіли	не захворіли
4	192	34	111

10.3. У таблиці 10.8 подано результати про фактичні обсяги збуту продукції (в тис. грн.) в п’яти районах міста.

Таблиця 10.8

Район	*1*	*2*	*3*	*4*	*5*
Фактичний обсяг збуту	110	130	70	90	100

Чи узгоджуються ці результати з припущенням маркетологів про те, що збут продукції в цих районах повинен бути однаковим?

10.4. На іспиті студент відповідає тільки на три питання з дисципліни (дисципліна поділена на 3 частини). Аналіз питань, заданих 60 студентам, показав, що 23 студенти отримали питання з першої, 15 – з другої і 22 – з третьої частини курсу. Чи можна вважати, що студент з рівною ймовірністю отримає питання з кожної із трьох частин дисципліни?

10.5. Знайти коефіцієнт рангової кореляції між урожайністю пшениці та картоплі на сусідніх полях за даними таблиці 10.9.

Таблиця 10.9

Роки	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*
Пшениця, (ц)	20,1	23,6	26,3	19,9	16,7	23,2	31,4	33,5	28,2	35,3	29,3	30,5
Картопля, (ц)	7,2	7,1	7,4	6,1	6,0	7,3	9,4	9,2	8,8	10,4	8,0	9,7

10.6. На основі даних дослідження умов життя сімей (таблиця 10.10) визначити взаємозв’язок між кількістю дітей та матеріальним достатком сімей за допомогою рангових коефіцієнтів кореляції.

Таблиця 10.10

Кількість дітей	Дохід		Разом
	Нижче прожиткового мінімуму	Достатній	
Бездітні та сім’ї з 1 дитиною	22	88	110
Сім’ї з 2 і більше дітьми	42	48	90
Разом	64	136	200

Знайти коефіцієнт кореляції Пірсона та перевірити його значущість, порівняти отримані результати.

Для завдань 10.7-10.14 необхідно перевірити вибірки на нормальність розподілу і залежно від результату провести відповідний аналіз (параметричний чи непараметричний).

10.7. Щоб підвищити обсяг продажів, фірма, яка торгує сиром, через мережу магазинів вирішила провести спеціальну рекламну акцію. Наведені нижче дані (табл. 10.11) відображають обсяг продажів (у тис. грн.) по днях, протягом яких рекламна акція проводилася (верхній рядок таблиці), і по днях, в які вона не проводилася (нижній рядок таблиці).

Таблиця 10.11

З рекламою	18	21	23	15	19	26	17	18	22	20	18	21	27
Без реклами	22	17	15	23	25	20	26	24	16	17	23	21	

Визначити, чи вплинула рекламна акція на підвищення обсягу продаж.

10.8. Викладач вирішив визначити, швидше чи повільніше його здібні студенти здають письмові тести: швидше тому, що вони швидше згадують засвоєні навички або повільніше тому, що на запис усього, що вони знають, іде більше часу. Зокрема, при вирішенні завдань зі статистики він записав отримані студентами оцінки в порядку здачі їх робіт (табл. 10.12).

Таблиця 10.12

Порядок здачі робіт	*Оцінки (бали)*									
1-10	94	70	85	89	92	98	63	88	74	85
11-20	69	90	57	86	79	72	80	93	66	74
21-30	50	55	47	59	68	63	89	51	90	88

Студентів, які набрали 90 і більше балів викладач вважає найбільш здібними студентами. Чи може він при рівні значущості 5% вважати, що здача робіт цими студентами носила випадковий характер?

Чи можна вважати, що студенти, які набрали 60 або більше балів, які вважаються такими, що пройшли тест, здали свої роботи у випадковій послідовності, на відміну від тих, хто не пройшов тест? (рівень значущості 5%).

10.9. Успішність студентів чотирьох груп оцінюється за 100-бальною шкалою. Оцінки студентів наведені в таблиці 10.13. Чи можна вважати, що медіани оцінок студентів по групах справді різні?

Таблиця 10.13

Група			
1	**2**	**3**	**4**
Бали			
77	44	26	7
31	78	70	28
59	38	55	19
48	20	61	39
40	25	73	55
59	29	61	36
57	51	63	19
22	74	79	11
13	54	50	9
16	56	45	80
22	47	33	92
5	40	45	10

10.10.Студентка відвідала кілька магазинів, щоб визначити, чи справді ціни на молоко значно відрізняються залежно від торгової марки. Її спостереження наводяться в таблиці 10.14. Чи можна зробити висновок, що ціни на молоко справді залежать від торгової марки?

Таблиця 10.14

Торгова марка A	Торгова марка B	Торгова марка C	Торгова марка D
Ціна (в умовних одиницях)			
61	52	47	67
55	58	52	63
57	54	49	68
60	55	49	69
58	57		65
62			

10.11. Час (у секундах) написання контрольних завдань одинадцятьма учнями до і після спеціальних вправ з усного рахунку наведений у таблиці 10.15. Чи можна вважати, що ці вправи поліпшили здібності учнів у розв'язанні задач?

Таблиця 10.15

До вправи	87	61	98	90	93	74	83	72	81	75	83
Після вправи	50	45	79	90	88	65	52	79	84	61	52

10.12. Кіноплівка чотирьох видів була надана трьом експертам для визначення кращої з них. Кожному експерту запропонували впорядкувати плівки за ступенем переваги. Бали (ранги), проставлені експертами, наведені в таблиці 10.16. Найбільший бал відповідає плівці найкращої якості.

Таблиця 10.16

Вид плівки / Експерти	1	2	3	4
1	2	1	3	4
2	2	1	4	3
3	2	1	4	3

Потрібно визначити, чи розрізняються види плівок і чи погоджені оцінки експертів.

10.13. Передбачається, що один із двох приладів, які визначають швидкість автомобіля, має систематичну помилку. Для перевірки цього припущення визначили швидкість десяти автомобілів, причому швидкість кожного фіксувалася одночасно двома приладами (таблиця 10.17).

Таблиця 10.17

1 прилад	70	85	63	54	65	80	75	95	52	55
2 прилад	72	86	62	55	63	80	78	90	53	57

Чи дозволяють ці результати стверджувати, що другий прилад дає завищені значення швидкості?

10.14. Під час презентації чотирьох нових торгових марок морозива п'ятнадцятьом покупцям було запропоновано спробувати всі торгової марки морозива і висловити своє ставлення до кожної марки в такому вигляді: 0 – подобається, 1 – не подобається. Відповіді покупців записані в таблиці 10.18. Потрібно перевірити гіпотезу: всі торгові марки морозива подобаються покупцям однаково.

Таблиця 10.18

Покупці	Торгова марка морозива			
	1	2	3	4
1	0	0	0	0
2	1	1	0	0
3	1	1	0	0
4	0	0	0	0
5	1	0	1	1
6	1	1	0	0
7	1	1	0	1
8	1	0	0	1
9	0	1	1	0
10	1	0	0	0
11	1	1	0	1
12	0	1	1	1
13	1	0	0	0
14	1	0	0	1
15	1	1	1	1

Лабораторна робота № 11
Аналіз динамічних рядів у системі STATISTICA

1. Основні теоретичні відомості про модуль Time Series/Forecasting

Однією з найбільш важливих задач статистики є аналіз динаміки, або зміни в часі досліджуваних явищ. Така задача розв'язується за допомогою так званих рядів динаміки (або часових, хронологічних рядів).

Рядом динаміки називається визначена на послідовності конкретних моментів (дат) або інтервалів (періодів) часу $\{t_1,...,t_n\}$ відповідна послідовність числових значень $\{y_1,...,y_n\}$ певного статистичного показника.

У загальному можна виділити такі етапи аналізу динамічних рядів:

1. Графічне подання динамічного ряду, вивчення зміни рівнів ряду за допомогою показників зміни рівнів і середніх показників динамічного ряду.
2. Виявлення основної тенденції (тренду) в рядах динаміки.
3. Вимірювання коливань у рядах динаміки (циклічні, сезонні та випадкові).
4. Дослідження наявності автокореляції між рівнями ряду динаміки та між залишковими величинами.
5. Прогнозування на основі моделей динамічних рядів.

У системі **STATISTICA 12** всі ці етапи реалізовані у модулі ***Time Series/Forecasting (Часові ряди та прогнозування)***.

Для активації модуля у вкладці ***Statistics*** групи ***Advanced/Multivariate (Додатковий/Багатовимірний аналіз)*** потрібно вибрати ***Advanced Models (Додаткові моделі)→Time Series/Forecasting*** або в меню ***Statistics*** – послідовність команд ***Advanced Linear/Nonlinear Models*** (***Додаткові лінійні/нелінійні моделі***)→***Time Series/Forecasting*** (рис. 11.1). Відкриється стартове вікно модуля (рис. 11.2).

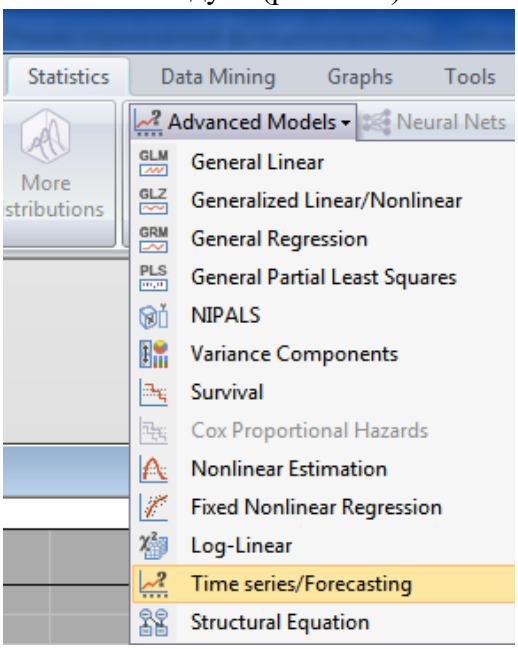


Рис. 11.1. Відкриття модуля ***Time Series/Forecasting*** у групі ***Advanced/Multivariate***

За допомогою кнопки ***Variables*** потрібно вибрати ім'я досліджуваної змінної (або змінних). Після вибору змінної (змінних) в інформаційній частині діалогового вікна в полі ***Variable*** з'явиться ім'я змінної (змінних), а в полі ***Long variable (series) name*** – довге ім'я змінної (змінних).

Весь подальший діалог дослідження динамічних рядів відбувається саме з цими вибраними змінними, які можна перетворювати, аналізувати, але не можна видаляти з поточного аналізу. У процесі роботи для вибору найбільш відповідного перетворення ряди багаторазово перетворюються, і щоб не зберігати зайву інформацію (невдалі перетворення), їх треба видаляти. Для цього служить кнопка ***Delete highlighted variable (Видалити підсвічені змінні)*** (рис. 11.2). Якщо для проведення подальших досліджень (можливо і в інших модулях системи **STATISTICA**) необхідно зберегти деякі перетворення, треба скористатися кнопкою ***Save variables (Зберегти змінні)*** (рис. 11.2).

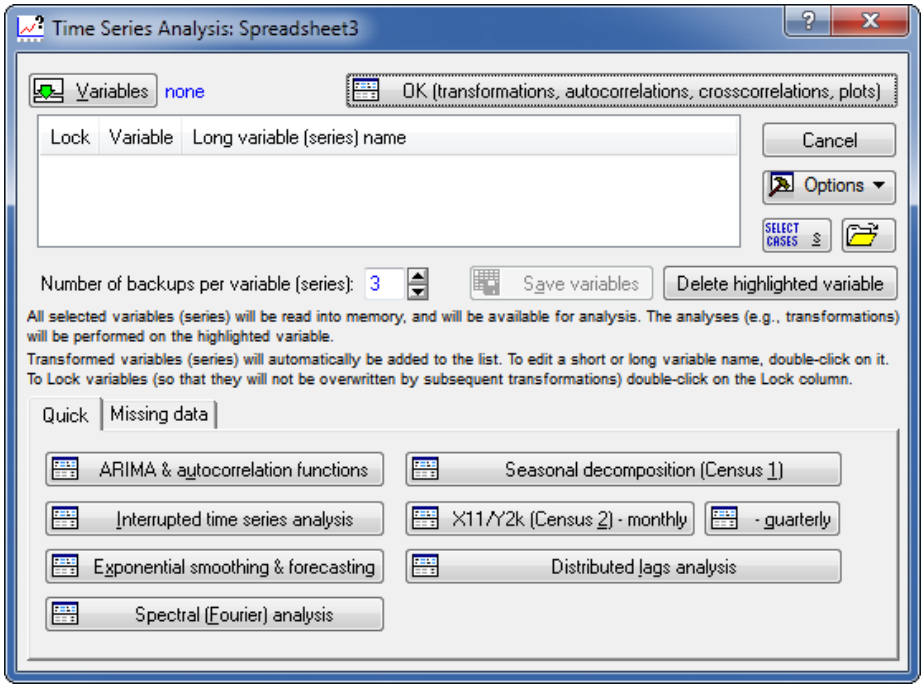


Рис. 11.2. Діалогове вікно модуля ***Time Series Analysis***

У полі ***Number of backups per variable (Число резервних копій для змінних)*** (рис. 11.2) визначають число перетворень поточного діалогу в інформаційній частині вікна. Якщо число перетворень перевищить указане число, то система зробить запит, чи зберегти чергове перетворення. Кнопка ***Select cases (Вибрати спостереження)*** призначена для вибору спостережень для аналізу. Кнопка ***OK (transformations, autocorrelations, crosscorrelations, plots) (Трансформація, автокореляція, кроскореляція, графіки)*** відкриває спеціальне вікно для перетворення ряду.

Вкладка ***Quick*** діалогового вікна ***Time Series Analysis*** містить кнопки, які визначають різні методи (процедури) аналізу часових рядів:

- ***ARIMA & autocorrelation functions (Модель авторегресії і проінтегрованого ковзного середнього та автокореляційні функції)***;
- ***Interrupted time series analysis (Аналіз перерваних часових рядів або моделі проінтегрованого ковзного середнього з інтервенцією)***;
- ***Exponential smoothing & forecasting (Експоненціальне згладжування і прогнозування)***, ***Seasonal decomposition (Census 1) (Сезонна декомпозиція***);
- ***Spectral (Fourier) analysis (Спектральний аналіз (Фур'є))***;
- ***X11/Y2k (Census 2)–montly (12-місячне сезонне корегування)***;
- ***quartenly (Квартальне сезонне корегування)***;
- ***Distributed Lags Analysis (Аналіз розподілених лагів)***.

На вкладці ***Missing Data (Пропущені дані)*** діалогового вікна ***Time Series Analysis*** (рис. 11.3) система пропонує різні можливості для заповнення пропущених значень:

- ***Overall mean*** – загальне середнє;
- ***Interpolation from adjacent points*** – інтерполяція за сусідніми точками (значеннями);
- ***Mean of N adjacent points*** – середнє за сусідніми точками (значеннями);
- ***Median of N adjacent points*** – медіана сусідніх точок (значень);
- ***Predicted values from linear trend regression*** – прогнозовані значення лінійного тренду.

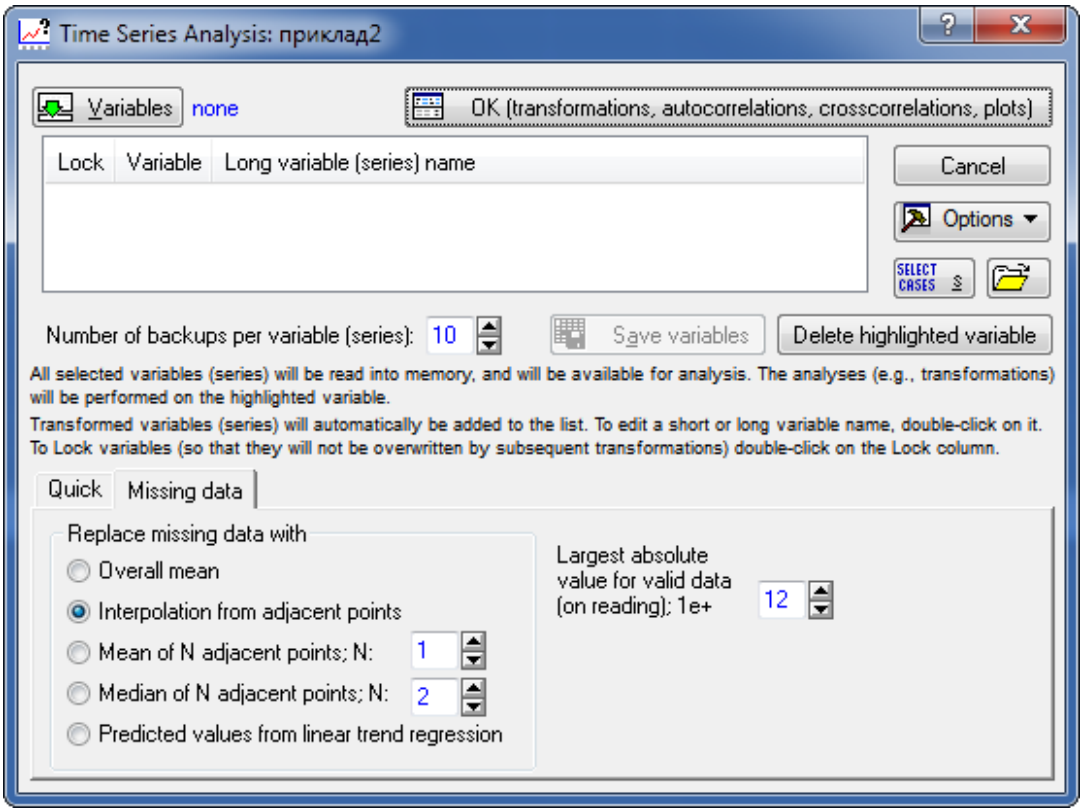


Рис. 11.3. Вкладка ***Missing Data*** діалогового вікна ***Time Series Analysis***

2. Виявлення основної тенденції (тренду) в динамічних рядах

Розглянемо основні етапи аналітичного вирівнювання динамічних рядів. Спочатку побудуємо графік початкових даних (фактичних рівнів динамічного ряду). Для цього у стартовому вікні модуля ***Time Series Analysis*** необхідно натиснути на кнопку ***OK (transformations, autocorrelations, crosscorrelations, plots)*** (див. рис. 11.2), в результаті чого на екрані з'явиться вікно ***Transformations of variables (Перетворення змінних)*** (рис. 11.4). У верхній частині вікна (рис. 11.4) записується ім'я динамічного ряду та його перетворення.

Далі перейдемо на вкладку ***Review & plot (Перегляд і графіки)***, яка містить такі кнопки:

- ***Review highlighted variables*** – перегляд підсвічених змінних;
- ***Review multiple variables*** – перегляд кількох змінних;
- ***Plot*** – побудова графіка;
- ***Plot two var list with different scales*** – побудова графіків змінних з двох списків у різних шкалах;
- ***Plot variables (series) after each transformation*** – побудова графіка змінних (ряду) після кожного перетворення в даному вікні;

- ***Display/plot subset only (from, through)*** – показ на екрані/побудова графіка для певної підмножини спостережень (з або через).

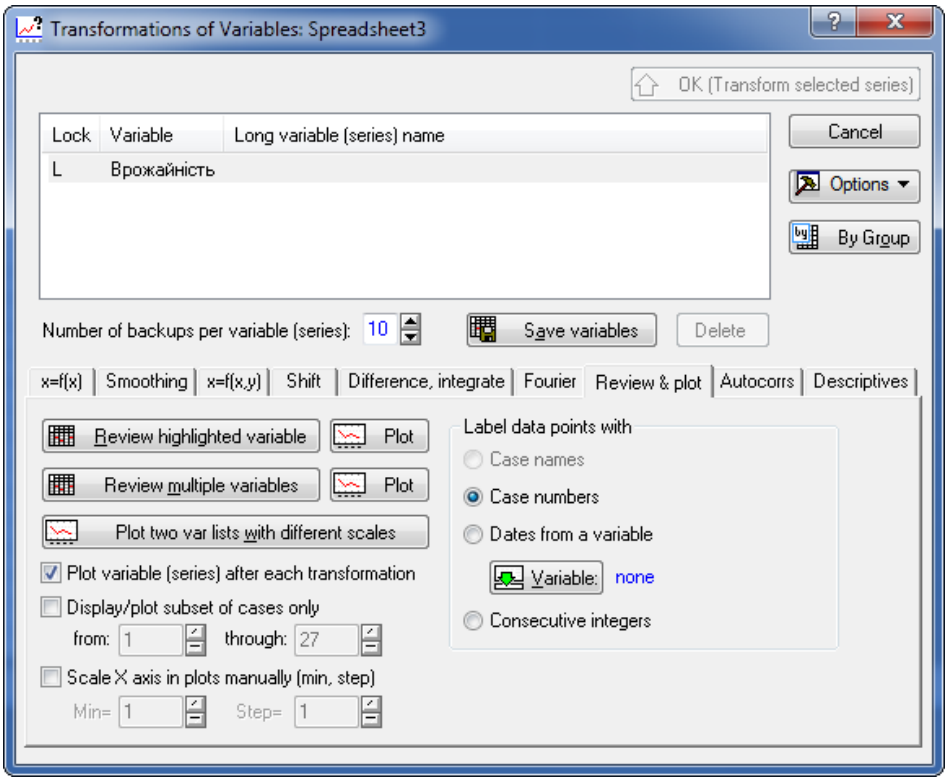


Рис. 11.4. Діалогове вікно ***Transformations of variables***

У правій частині вікна (рис. 11.4) в рамці ***Label data points with (Мітки точок даних)*** розташовані опції позначень спостережень на графіку (***Case names*** – імена спостережень, ***Case numbers*** – номери спостережень, ***Dates from a var*** – значення змінної, ім'я якої вказується за допомогою кнопки ***Variable***, ***Consecutive integers –*** послідовність цілих чисел).

Перед побудовою графіків необхідно задати потрібні позначення шкал X і Y (вісі абсцис та ординат). Для цього треба вказати потрібні значення мінімуму (номер спостереження, з якого будується графік) і кроку шкали X (абсцис) в опції ***Scale X axis in plots manually (min, step)*** (ліва нижня частина вікна), а також вибрати мітки для шкали Y (ординат) в опціях ***Label data points with***.

За допомогою кнопки ***Plot*** (поряд з кнопкою ***Review highlighted variables***) можна побудувати графік динамічного ряду, який дозволить провести попередню візуальну оцінку даних ряду (рис. 11.4).

Для виявлення основної тенденції зміни рівня ряду (тренду) методом ковзних середніх необхідно в цьому ж вікні ***Transformations of Variables*** перейти на вкладку ***Smoothing (Згладжування)*** та виділити опцію ***N-pts mov. averg. (Згладжування за методом ковзних середніх)***. Потім необхідно задати ***N*** (інтервал усереднення, тобто кількість точок, за допомогою яких визначається середня для рівня y_t) і натиснути кнопку ***OK*** (***Transform selected series***) ***(Трансформація вибраних рядів)***. На екрані з'явиться графік згладженого ряду. Щоб переглянути одночасно початкові дані (фактичні рівні ряду) і результати згладжування, за допомогою простого ковзного середнього потрібно знову перейти на вкладку ***Review & plot*** і натиснути кнопку ***Review multiple variables*** (рис. 11.5). Якщо натиснути кнопку ***Plot*** поряд з цією кнопкою, то можна переглянути ковзні середні на графіку одночасно з початковими фактичними рівнями ряду (рис. 11.6).

На вкладці $x = f(x)$ діалогового вікна ***Transformations of Variables*** (рис. 11.4) є певні опції, що дозволяють виконати різні перетворення початкового часового ряду: додавання константи, піднесення до цілого та дробового степеня, логарифмування, стандартизація даних тощо. Перетворення можна виконувати послідовно. Одна з цілей перетворень полягає в тому, щоб привести ряд до стаціонарності. У випадку лінійного тренду це можна зробити за допомогою опції ***Trend subtract (*** $x = x - (a + b * t)$***) (Видалення тренду (*** $x = x - (a + b * t)$***))***. Параметри a і b лінійного тренду $a + bt$ можуть задаватися або оцінюватися за початковими даними (фактичними рівнями ряду).

Для побудови корелограми треба перейти на вкладку ***Autocorrs (Автокореляція)*** діалогового вікна ***Transformations of Variables*** і натиснути кнопку ***Autocorrelations***. Система **STATISTICA** побудує корелограму й таблицю з коефіцієнтами кореляції. Якщо автокореляційна функція не має тенденції до затухання, можна говорити про нестаціонарність ряду.

Case	Listing of selected variables (ser	
	Врожайність	Врожайність trns.
1	1920,000	
2	2020,000	
3	2060,000	1995,000
4	1960,000	2015,000
5	1960,000	2020,000
6	2140,000	2007,500
7	1980,000	1983,750
8	1940,000	1976,250
9	1790,000	2043,750
10	2250,000	2137,500
11	2410,000	2228,750
12	2260,000	2286,250
13	2200,000	2252,500
14	2300,000	2211,500
15	2090,000	2230,500
16	2252,000	2253,000
17	2360,000	2274,250
18	2320,000	2274,000
19	2240,000	2247,000
20	2100,000	2230,125
21	2296,000	2231,375
22	2249,000	2275,000
23	2321,000	2297,125
24	2368,000	2287,250
25	2205,000	2298,625
26	2261,000	
27	2400,000	

Рис. 11.5. Фактичні рівні та середні ковзні динамічного ряду

Для дослідження та прогнозування рядів динаміки, в яких варіація рівнів може мати сезонний характер, здійснюють декомпозицію динамічного ряду. Декомпозиція динамічного ряду дозволяє виділити компоненти динамічного ряду, а саме: тренд (загальну тенденцію), сезонну, циклічну та випадкові складові. Найпоширенішими моделями декомпозиції ряду є моделі з адитивною та мультиплікативною компонентами.

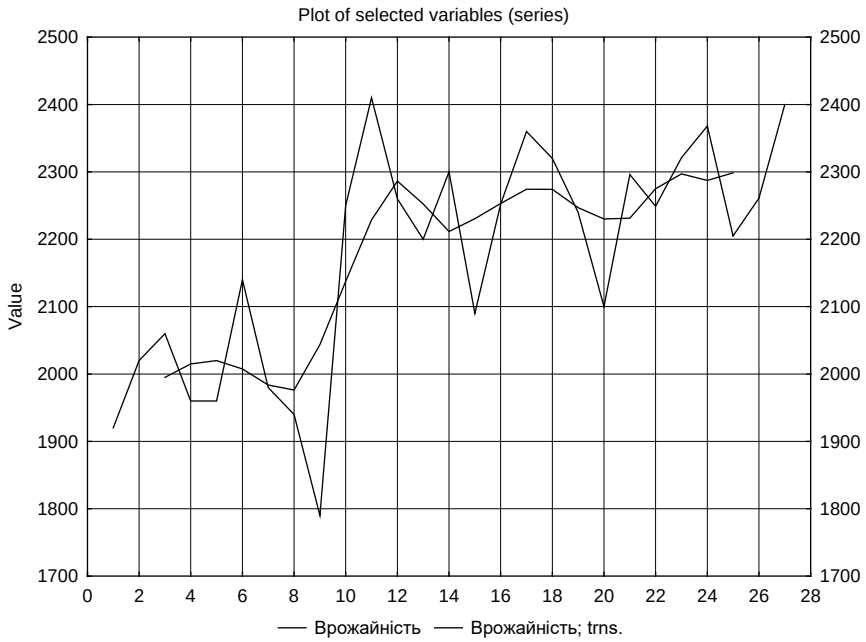


Рис. 11.6. Динаміка врожайності та вирівнювання ряду за допомогою методу середньої ковзної

Щоб виділити компоненти динамічного ряду, необхідно натиснути кнопку ***Seasonal decomposition (Census 1) (Сезонна декомпозиція)*** у діалоговому вікні модуля ***Time Series/Forecasting*** (рис. 11.2) (Для повернення до діалогового вікна ***Time Series Analysis*** потрібно натиснути кнопку ***Cancel (Скасувати)*** у вікні ***Transformations of variables***). У результаті відкриється діалогове вікно ***Ratios-to-Moving Averages Classical Seasonal Decomposition (Census Method I) (Коефіцієнти ковзних середніх, класична сезонна декомпозиція (метод Census I))*** (рис. 11.7).

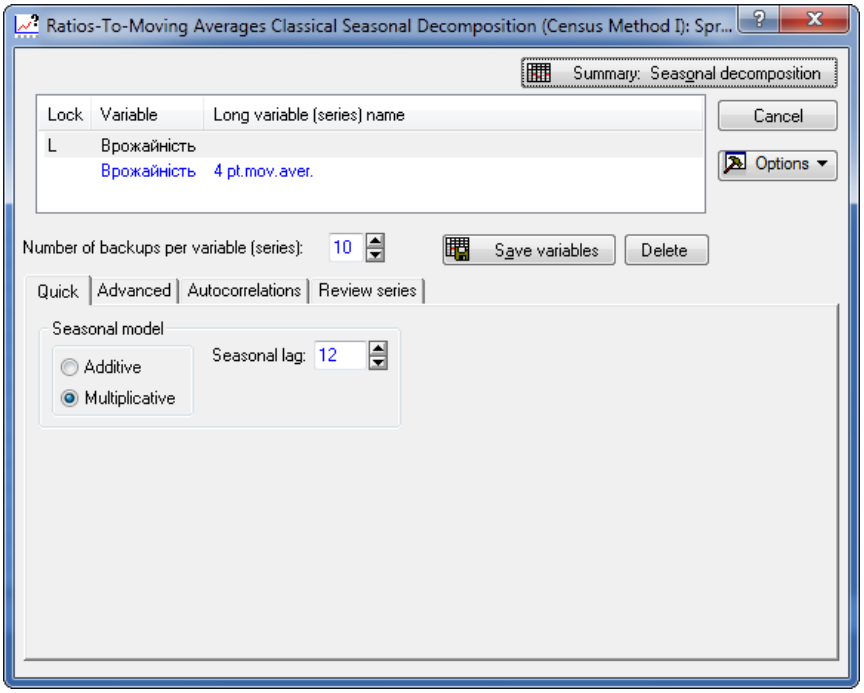


Рис. 11.7. Діалогове вікно ***Ratios-to-Moving Averages Classical Seasonal Decomposition (Census Method I)***

На вкладці ***Quick*** діалогового вікна ***Ratios-to-Moving Averages Classical Seasonal Decomposition (Census Method I)*** розміщені опції, що дозволяють задати модель декомпозиції, у полі ***Seasonal model (Сезонна модель)***:

- ***Additive (Адитивна)***;
- ***Multiplicative (Мультиплікативна)***.

У полі ***Seasonal lag (Сезонний лаг)*** задається довжина сезонного періоду.

На вкладці ***Advanced*** діалогового вікна ***Ratios-to-Moving Averages Classical Seasonal Decomposition (Census Method I)*** можна задати опцію ***Centered moving averages (for even Seasonal lag only) (Центровані ковзні середні для парного сезонного лага)***, що дозволяє користувачеві при парному інтервалі усереднення вибрати одну з двох можливостей: розрахувати ковзне середнє з однаковими вагами або ж так, щоб перше і останнє спостереження у вікні мали різні ваги. Другий метод використовується, якщо встановлений прапорець. Якщо ж довжина сезонного періоду непарна, то вибір цієї опції не впливає на обчислення. Наступна група опцій вкладки ***Advanced On OK append components to active work area (Додавання компонентів у активній робочій області)*** дозволяє додати до активного робочого простору такі складові: ***Moving averages (Ковзні середні)***, ***Ratios/Differences (Відношення/Різниці)***, ***Seasonal factors (Сезонні фактори)***, ***Seasonal adj. series (Ряд, скорегований на сезонну складову)***, ***Smoothed trend cycle (Згладжену тренд-циклічну компоненту)*** та ***Irregular components (Нерегулярну складову)***.

Для здійснення сезонної декомпозиції треба вибрати ***Seasonal factors, Smoothed trend cycle*** та ***Irregular components*** у вкладці ***Advanced*** і натиснути кнопку ***Summary: Seasonal decomposition (Результати: Сезонна декомпозиція)***. Результати сезонної декомпозиції динамічного ряду виводиться у вигляді таблиці (рис. 11.8).

Seasonal Decomposition: Additive season (4) (Spreadsheet3)
Врожайність

Case	Врожайність	Moving Averages	Diffrncs	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	1920,000			-48,9479	1968,948	1986,369	-17,421
2	2020,000			55,0521	1964,948	1990,865	-25,917
3	2060,000	1990,000	70,000	21,3021	2038,698	1999,855	38,843
4	1960,000	2000,000	-40,000	-27,4063	1987,406	2011,934	-24,528
5	1960,000	2030,000	-70,000	-48,9479	2008,948	2018,772	-9,824
6	2140,000	2010,000	130,000	55,0521	2084,948	2016,105	68,843
7	1980,000	2005,000	-25,000	21,3021	1958,698	1980,966	-22,269
8	1940,000	1962,500	-22,500	-27,4063	1967,406	1975,267	-7,861
9	1790,000	1990,000	-200,000	-48,9479	1838,948	2020,994	-182,046
10	2250,000	2097,500	152,500	55,0521	2194,948	2143,883	51,065
11	2410,000	2177,500	232,500	21,3021	2388,698	2246,522	142,176
12	2260,000	2280,000	-20,000	-27,4063	2287,406	2286,378	1,028
13	2200,000	2292,500	-92,500	-48,9479	2248,948	2252,105	-3,157
14	2300,000	2212,500	87,500	55,0521	2244,948	2215,216	29,731
15	2090,000	2210,500	-120,500	21,3021	2068,698	2212,522	-143,824
16	2252,000	2250,500	1,500	-27,4063	2279,406	2255,934	23,472
17	2360,000	2255,500	104,500	-48,9479	2408,948	2289,216	119,731
18	2320,000	2293,000	27,000	55,0521	2264,948	2272,994	-8,046
19	2240,000	2255,000	-15,000	21,3021	2218,698	2243,855	-25,157
20	2100,000	2239,000	-139,000	-27,4063	2127,406	2218,712	-91,306
21	2296,000	2221,250	74,750	-48,9479	2344,948	2243,994	100,954
22	2249,000	2241,500	7,500	55,0521	2193,948	2265,994	-72,046
23	2321,000	2308,500	12,500	21,3021	2299,698	2297,411	2,287
24	2368,000	2285,750	82,250	-27,4063	2395,406	2299,267	96,139
25	2205,000	2288,750	-83,750	-48,9479	2253,948	2293,661	-39,713
26	2261,000	2308,500	-47,500	55,0521	2205,948	2279,531	-73,583
27	2400,000			21,3021	2378,698	2272,466	106,231

Рис. 11.8. Результати сезонної декомпозиції динамічного ряду

У таблиці результатів сезонної декомпозиції наведено такі дані:

- ***Moving Averages*** – прості ковзні середні;
- ***Diffrncs –*** різниці між фактичними рівнями ряду та ковзними середніми;
- ***Seasonal factors*** – сезонні компоненти;
- ***Adjusted Series*** – скорегований ряд динаміки (десезоналізовані рівні);
- ***Smoothed Trend–c. –*** ряд, скорегований на згладжену тренд-циклічну компоненту;
- ***Irreg. Compon. –*** нерегулярна складова ряду динаміки.

Графічно подати всі компоненти динамічного ряду можна за допомогою вкладки ***Review series***, вибравши кнопку ***Plot*** поряд з кнопкою ***Review multiple variables***.

3. Типовий приклад

Кількість відвідувачів за місяць кінотеатру подана у таблиці 11.1

Таблиця 11.1

Дні місяця	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
Кількість відвідувачів (тис.)	*1,65*	*2,59*	*6,18*	*6,26*	*6,44*	*7,16*	*10,56*	*10,93*	*9,53*	*10,64*	*17,43*	*14,72*	*15,5*

Дні місяця	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*
Кількість відвідувачів (тис.)	*15,01*	*17,83*	*18,43*	*17,69*	*19,8*	*22,64*	*22,86*	*21,56*	*22,16*	*25,82*	*26,5*

Дослідити ряд динаміки та встановити наявність сезонних коливань.

Розв'язування. *Для аналізу динамічного ряду використаємо модуль* ***Time Series/Forecasting.*** *Спочатку побудуємо графік динамічного ряду. Для цього перейдемо в діалогове вікно* ***Transformations of variables*** *і на вкладці* ***Review & plot*** *натиснемо на кнопку* ***Plot*** *поряд з* ***Review highlighted variables.*** *У результаті отримаємо графік (рис. 11.9).*

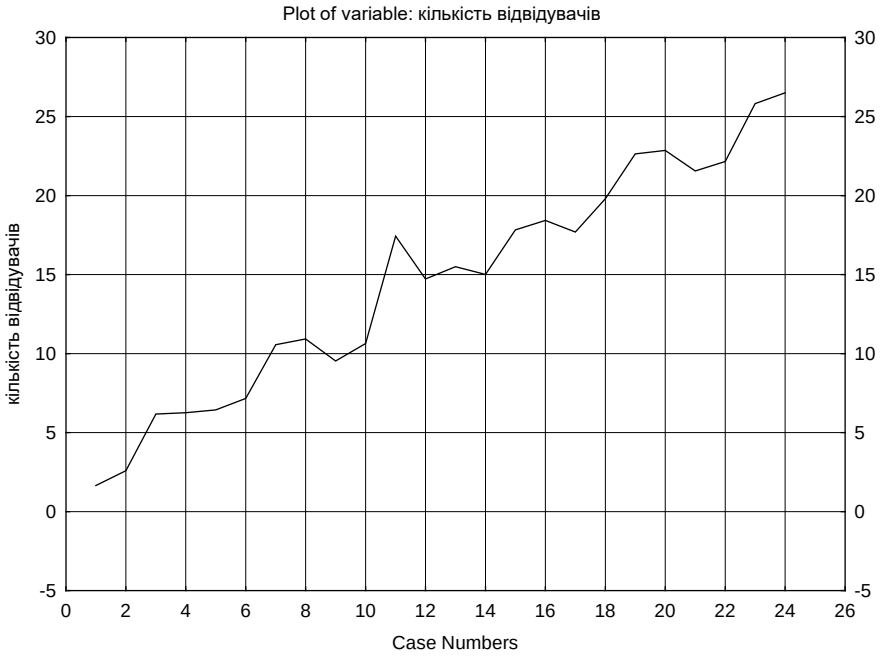


Рис. 11.9. Графік динамічного ряду "Кількість відвідувачів"

Як бачимо, графік не дає відповіді про наявність сезонних коливань, однак показує тенденцію до зростання. Щоб виявити, чи присутня сезонність у даному ряду, перетворимо його за допомогою опції ***Trend subtract*** *(*$x = x - (a + b * t)$*). Для цього перейдемо на вкладку* $x = f(x)$*, позначимо дану опцію і натиснемо на кнопку* ***OK*** *(****Transform selected series****). У результаті система побудує графік, поданий на рис. 11.10.*

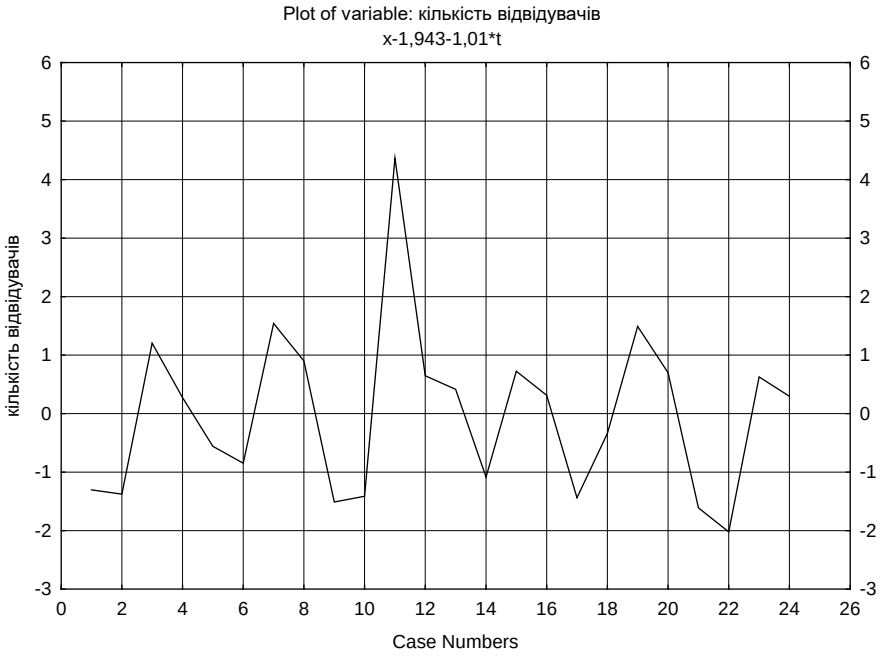


Рис. 11.10. Скорегований ряд без тренду

На даному графіку чітко видно сезонну складову з розміром інтервалу 4.

Щоб визначити сезонні компоненти, скористаємося аналізом ***Seasonal decomposition****, виберемо вид моделі – адитивна, вкажемо розмір сезонного лага – 4 та натиснемо кнопку* ***Summary: Seasonal decomposition (Результати: Сезонна декомпозиція).*** *Результати розрахунку подані в таблиці (рис. 11.11).*

Seasonal Decomposition: Additive season (4); Centered means (Spreadsheet15)
кількість відвідувачів: x-1,943-1,01*t

Case	кількість відвідувачів trnsfrmd	Moving Averages	Diffrncs	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	-1,30430			-1,01012	-0,29417	-0,382697	0,088522
2	-1,37512			-1,19712	-0,17800	-0,336413	0,158417
3	1,20406	-0,207193	1,41125	1,74112	-0,53707	-0,243846	-0,293222
4	0,27323	-0,048015	0,32125	0,46613	-0,19289	-0,064112	-0,128778
5	-0,55759	0,059913	-0,61750	-1,01012	0,45254	0,103538	0,349000
6	-0,84841	0,180341	-1,02875	-1,19712	0,34872	0,199050	0,149667
7	1,54077	0,139520	1,40125	1,74112	-0,20036	0,101756	-0,302111
8	0,89995	-0,050052	0,95000	0,46613	0,43382	0,003712	0,430111
9	-1,51087	0,232876	-1,74375	-1,01012	-0,50075	0,151362	-0,652111
10	-1,41170	0,554554	-1,96625	-1,19712	-0,21457	0,469096	-0,683667
11	4,36748	0,763733	3,60375	1,74112	2,62636	0,970691	1,655667
12	0,64666	1,045411	-0,39875	0,46613	0,18054	0,949314	-0,768778
13	0,41584	0,630839	-0,21500	-1,01012	1,42596	0,719186	0,706778
14	-1,08498	0,133767	-1,21875	-1,19712	0,11214	0,131365	-0,019222
15	0,72420	-0,139554	0,86375	1,74112	-1,01693	-0,237040	-0,779889
16	0,31337	-0,277876	0,59125	0,46613	-0,15275	-0,263973	0,111222
17	-1,43745	-0,088698	-1,34875	-1,01012	-0,42732	-0,126323	-0,301000
18	-0,33827	0,055480	-0,39375	-1,19712	0,85886	0,144744	0,714111
19	1,49091	0,082159	1,40875	1,74112	-0,25022	0,045228	-0,295444
20	0,70009	-0,149913	0,85000	0,46613	0,23396	-0,107260	0,341222
21	-1,61073	-0,468235	-1,14250	-1,01012	-0,60061	-0,482943	-0,117667
22	-2,02156	-0,626557	-1,39500	-1,19712	-0,82443	-0,648543	-0,175889
23	0,62762			1,74112	-1,11350	-0,702420	-0,411083
24	0,29680			0,46613	-0,16932	-0,729359	0,560034

Рис. 11.11. Результат декомпозиції динамічного ряду

Переглянути ковзні середні на графіку одночасно з початковими фактичними рівнями ряду можна натиснувши кнопку ***Plot*** *поряд з кнопкою* ***Review multiple variables*** *у вкладці* ***Review series*** *діалогового вікна* ***Ratios-to-Moving Averages Classical Seasonal Decomposition (Census Method I).*** *З'явиться діалогове вікно вибору рядів для побудови графіків (рис. 11.12)* ***Select Variables for Plot/Spreadsheet (Вибрати змінні для***

графіку/таблиці). *Укажемо необхідні ряди і натиснемо* ***OK****. Графік динаміки кількості відвідувачів і вирівнювання ряду за допомогою методу середньої ковзної наведений на рис. 11.13.*

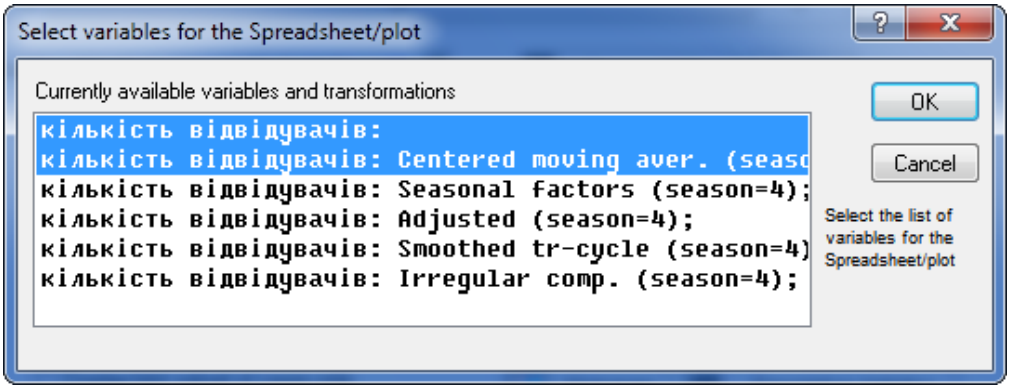


Рис. 11.12. Вибір рядів для побудови графіка

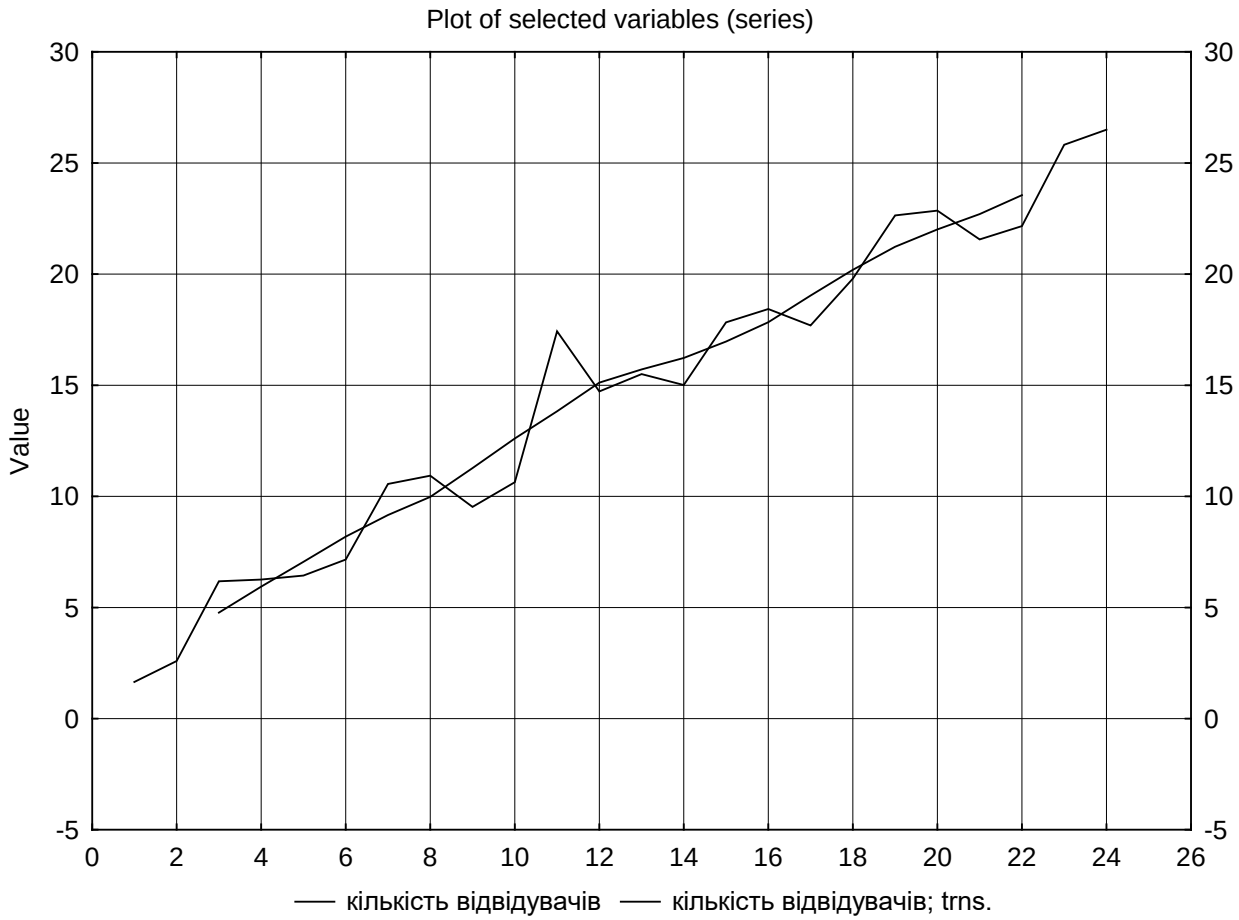


Рис. 11.13. Динаміка кількості відвідувачів і вирівнювання ряду за допомогою методу середньої ковзної

Аналогічно можна побудувати графік динамічного ряду з усіма його компонентами (рис. 11.14), вибравши всі ряди у діалоговому вікні ***Select Variables for Plot/Spreadsheet*** *(рис. 11.12).*

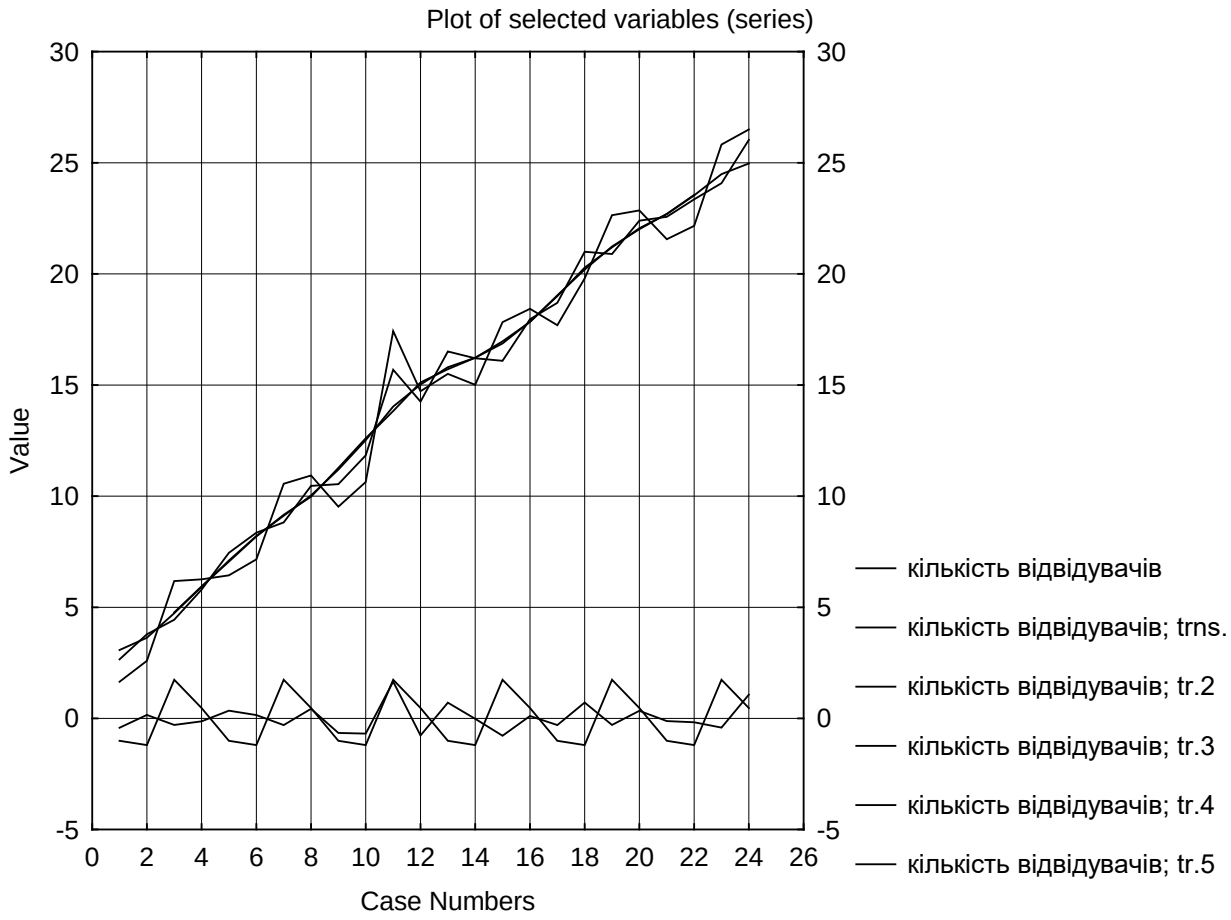


Рис. 11.14. Графік динамічного ряду зі всіма компонентами (кількість відвідувачів; tms. – динамічний ряд, вирівняний за допомогою ковзної (центрованої) середньої; кількість відвідувачів; tr.2 – сезонні компоненти; кількість відвідувачів; tr.3 – десезоналізований ряд; кількість відвідувачів; tr.4 – ряд, скорегований на згладжену тренд-циклічну компоненту; кількість відвідувачів; tr.5 – нерегулярна складова ряду динаміки)

Завдання для самостійної роботи

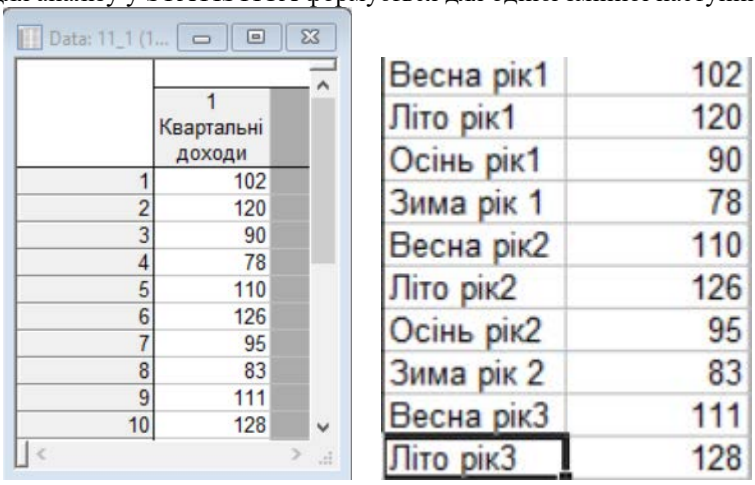
11.1. Квартальні доходи власника фірми з виробництва човнів аналізував квартальні доходи за останні 5 років (у тис. грн.) подані у табл. 11.2.

Таблиця 11.2

Рік	*Весна*	*Літо*	*Осінь*	*Зима*
1	102	120	90	78
2	110	126	95	83
3	111	128	97	86
4	115	135	103	91
5	122	144	110	98

Знайти 4-квартальні центровані ковзні середні, відношення (у відсотках) вихідних даних до ковзної середньої та сезонні індекси для кожного кварталу. Побудувати автокореляційну функцію.

Примітка. Файл для аналізу у STATISTICA формується для однієї змінної наступним чином.



11.2. Комісія визначила витрати енергії (табл. 11.3), виходячи з квартальних витрат натурального газу (в млн. $м^3$).

Таблиця 11.3

Рік	*Зима*	*Весна*	*Літо*	*Осінь*
1	293	246	231	282
2	301	252	227	291
3	304	259	239	296
4	306	265	240	300

Визначити сезонні індекси і виключити з фактичних рівнів динамічного ряду сезонну складову. Методом найменших квадратів знайти параметри лінійного тренду. Побудувати графіки фактичних рівнів ряду, рівнів без сезонної складової та без тренда.

Примітка. Файл для аналізу у STATISTICA формується для однієї змінної аналогічно 11.1.

11.3. Дані продажів місцевого виробника пива наведені у таблиці 11.4.

Таблиця 11.4

Рік	*Продажі за квартал, в тис. грн.*			
	I	*II*	*III*	*IV*
1	19	24	38	25
2	21	28	44	23
3	23	31	41	23
4	24	35	48	21

Обчислити сезонні індекси (використовувати центровані середні значення за 4 квартали). Виключити сезонну складову з фактичних даних. Методом найменших квадратів знайти параметри прямої, яка найкращим способом описує основну тенденцію динамічного ряду. Визначити циклічну компоненту в цьому динамічному ряді, виключивши тренд з фактичних рівнів динамічного ряду даних. Здійснити прогноз на наступні роки.

Примітка. Файл для аналізу у STATISTICA формується для однієї змінної аналогічно 11.1.

11.4. Кількість гостей гірського курорту протягом кожного сезону за останні 5 років подана у табл. 11.5. Обчислити сезонний індекс для кожного кварталу. Якщо влітку на курорті працювало 50 людей, скільки людей потрібно найняти взимку?

Таблиця 11.5

Рік	*Весна*	*Літо*	*Осінь*	*Зима*
1	200	300	125	325
2	175	250	150	375
3	225	300	200	450
4	200	350	225	375
5	175	300	200	350

Примітка. Файл для аналізу у STATISTICA формується для однієї змінної аналогічно 11.1.

11.5. Декан економічного факультету склав таблицю відвідування занять студентами за останні 5 років (табл. 11.6).

Таблиця 11.6

Рік	*Осінь*	*Зима*	*Весна*
1	220	203	193
2	235	208	206
3	236	206	209
4	241	215	206
5	239	221	213

Обчислити сезонні індекси. Виключити сезонну складову з фактичних рівнів ряду. Методом найменших квадратів знайти параметри тренда, який найліпше описує основну тенденцію часового ряду.

Примітка. Файл для аналізу у STATISTICA формується для однієї змінної аналогічно 11.1.

Лабораторна робота № 12
Кластерний аналіз

1. Основні теоретичні відомості про модуль Cluster Analysis

Кластерний аналіз – розбиття множини досліджуваних об'єктів і ознак на однорідні в деякому розумінні групи, або кластери.

У системі **STATISTICA** реалізовані класичні методи кластерного аналізу, включаючи ієрархічну кластеризацію, метод k –средніх, двовходова кластеризація.

Для запуску модуля ***Cluster Analysis (Кластерний аналіз)*** необхідно вибрати у вкладці ***Statistics*** у групі ***Advanced/Multivariate (Додатковий/Багатовимірний аналіз) Mult/Exploratory (Багатовимірні дослідницькі методи)→ Cluster*** або в меню ***Statistics*** – послідовність команд ***Multivariate Exploratory Techniques→Cluster Analysis***. Відкриється діалогове вікно ***Clustering Method (Методи кластеризації)*** (рис. 12.1).

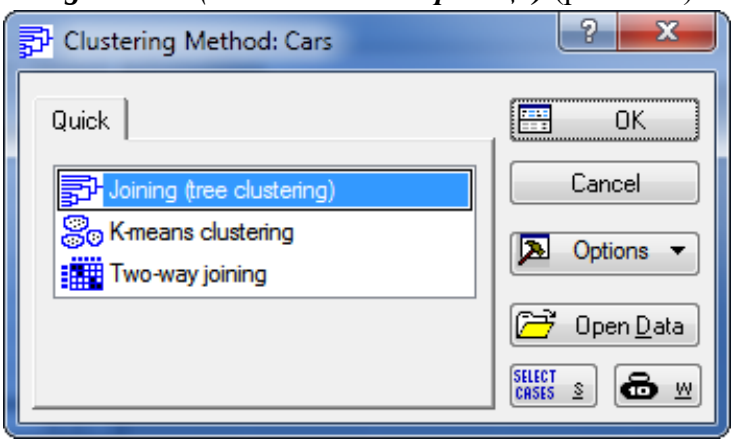


Рис. 12.1. Діалогове вікно ***Clustering Method***

На вкладці ***Quick*** знаходиться список методів кластерного аналізу, реалізованих у програмі **STATISTICA**:

- ***Joining tree clustering (Деревоподібна (ієрархічна) кластеризація)***;
- ***k–means clustering (Метод k -середніх)***;
- ***Two–way joining (Двовходова кластеризація).***

Вибравши потрібний метод, натиснемо ***OK*** і відкриємо діалогове вікно кластерного аналізу відповідного методу (рис. 12.2).

Вибір змінних для аналізу здійснюється через кнопку ***Variables***.

На вкладці ***Advanced*** діалогового вікна ***Cluster Analysis: K-Means Clustering*** у полі ***Cluster (Кластер)*** треба вибрати об'єкти для кластеризації: ***Variables (columns) (Змінні (стовпці))*** або ***Cases (rows)*** (***Спостереження (рядки)***); у полі ***Number of clasters (Число кластерів)*** потрібно вказати число кластерів; у полі ***Number of iterations (Число ітерацій)*** задається максимальне число ітерацій, що будуть використовуватися при побудові кластерів.

Якщо необхідно провести кластеризацію не по всіх об'єктах, треба скористатися кнопкою ***Select cases***.

Група опцій ***Initial cluster centers*** (рис. 12.1) дозволяє задати початкові центри кластерів:

- ***Choose observations to maximize initial between–cluster distances (Вибрати спостереження максимізувавши початкові відстані між кластерами)***;
- ***Sort distances and take observations at constant intervals (Сортувати відстані та вибрати спостереження на постійних інтервалах)***;

- ***Choose the first N (Number of clasters) observations (Вибрати перші N (число кластерів) спостережень).***

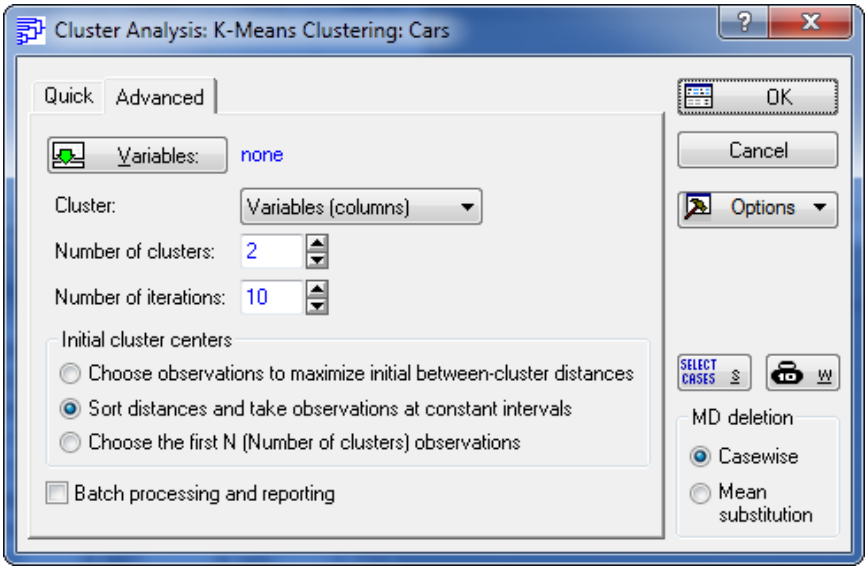


Рис. 12.2. Діалогове вікно кластерного аналізу методом *k* -середніх

2. Кластеризація за допомогою методу K-Means Clustering

Після вибору методу ***k-means clustering*** відкриється вікно результатів ***k-means Clustering Results (Результати кластерного аналізу методом k-means)*** (рис. 12.3)

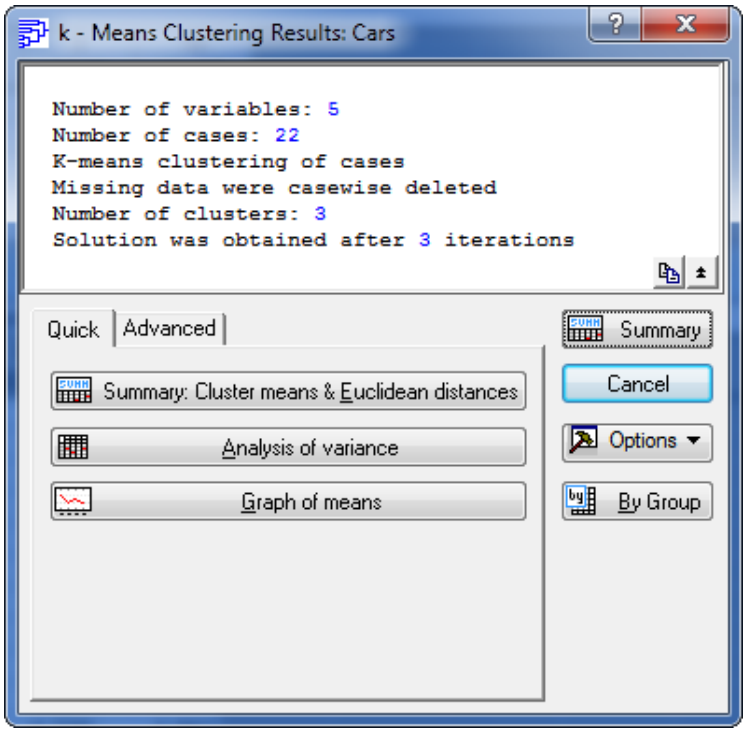


Рис. 12.3. Вікно результатів кластерного аналізу

У верхній інформаційній частині вікна виведена така інформація:

- ***Number of variables*** – кількість змінних;
- ***Number of cases*** – число спостережень;
- ***k–means clustering of cases*** – метод *k*-середніх;
- ***Missing data were casewise deleted*** – пропущені значення видалені;
- ***Number of clusters*** – число кластерів;

- ***Solution was obtained after * iterations*** – розв'язок знайдено після * ітерацій.

Функціональне призначення кнопок на вкладці ***Advanced*** діалогового вікна результатів ***k-means Clustering Results*** таке:

- ***Summary: Cluster means & Euclidean distances (Результат: середні кластерів та евклідові відстані)*** призначена для виведення таблиць, у першій з яких указані середні для кожного кластера (усереднення проводиться усередині кластера), у другій – евклідові відстані та квадрати евклідових відстаней між кластерами.
- ***Analysis of variance (Дисперсійний аналіз)*** виводить таблицю дисперсійного аналізу. У таблиці знайдено значення міжгрупових (*Between SS*) і внутрігрупових (*Within SS*) дисперсій ознак. Чим менше значення внутрігрупової дисперсії та більше значення міжгрупової дисперсії, тим краще ознака характеризує належність об'єктів до кластера і тим "якісніша" кластеризація. Параметри *F* і *p* також характеризують внесок ознаки в розподіл об'єктів на групи. Кращій кластеризації відповідають великі значення першого та менші значення другого параметра. Ознаки з великими значеннями *p* (більше 0,05) можна з процедури кластеризації виключити.
- ***Graph of means (Графіки середніх)*** дозволяє переглянути середні значення для кожного кластера на лінійному графіку.
- ***Descriptive statistics for each cluster (Описова статистика для кожного кластера)*** виводить таблицю з описовими статистиками (основними статистичними показниками) для кожного кластера.
- ***Members of each clusters & distances (Члени кожного кластеру та відстані)*** призначена для перегляду розподілу об'єктів по кластерах. У таблиці також буде вказано відстань від об'єкта до центру кластера.
- ***Save classifications and distances (Зберігати результати класифікації та відстані)*** зберігає результати класифікації у файлі **STATISTICA** для подальшого дослідження. При цьому в новому файлі кожному спостереженню програмою присвоюється номер кластера, до якого він був віднесений при класифікації.

3. Кластерний аналіз за допомогою двовходової кластеризації

Розглянемо процедуру одночасної кластеризації за змінними (стовпцями) і за спостереженнями (рядками) за допомогою методу ***Two-way joining*** діалогового вікна ***Clustering Method*** (рис. 12.1).

На вкладці ***Advanced*** діалогового вікна ***Cluster Analysis: Two-Way Joining*** є можливість вибрати пороговий параметр ***Threshold Value (Значення порогу)***. Пороговий параметр визначає, коли алгоритм розглядає в матриці даних два числа як однакові, а потім приписує їх до одного кластера. Якщо ця величина дуже велика (по відношенню до чисел у матриці даних), то буде сформований тільки один кластер; якщо вона дуже мала, то кластером буде кожна точка даних. Параметр може задати користувач – ***User defined***. Але для більшості випадків рекомендується величина за замовчуванням – ***Computed from data*** (Загальне середньоквадратичне відхилення, що ділиться на 2). Після натискання ***OK*** на екрані з'явиться вікно результатів (рис. 12.4). У верхній інформаційній частині вікна вказано число змінних; число спостережень; порогове значення; число отриманих блоків розбиття; середньоквадратичне відхилення.

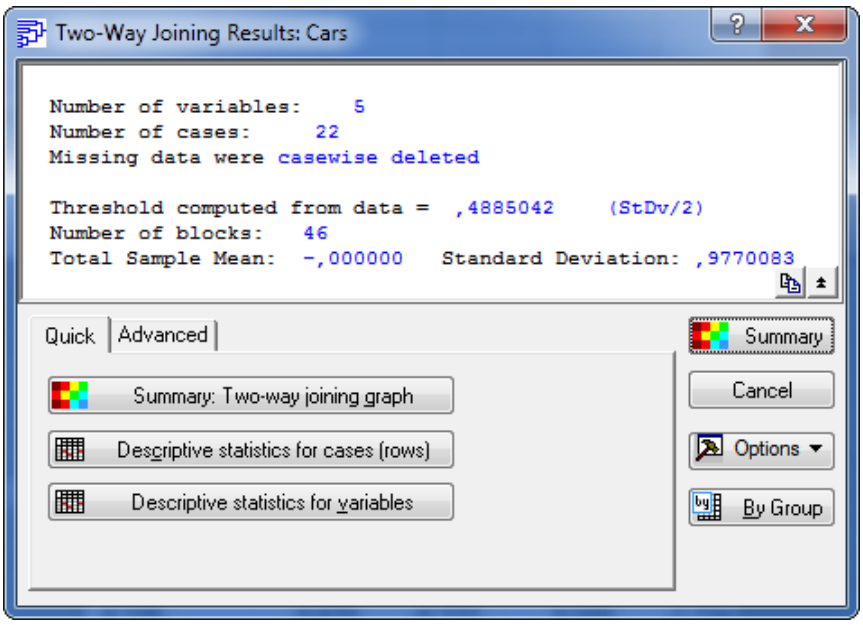


Рис. 12.4. Вікно результатів кластерного аналізу методом ***Two–way joining***

Вкладка ***Advanced*** діалогового вікна ***Two-Way Joining Results*** дозволяє переглянути:

- ***Summary: Two–way joining graph*** – графічне подання результатів.
- ***Descriptive statistics for cases (row)*** – описові статистики для спостережень.
- ***Descriptive statistics for variables*** – описові статистики для змінних.
- ***Reordered data matrix*** – невпорядкована матриця значень.

4. Ієрархічний метод кластеризації

Для кластеризації методом ***Joining tree clustering*** треба у діалоговому вікні ***Clustering Method*** обрати ***Joining tree clustering*** і натиснути на ***OK***. З'явиться діалогове вікно вибору змінних задання параметрів кластеризації (рис. 12.5). У списку ***Input file*** (Вхідні дані) можна вибрати вид файлу початкових даних: ***Raw data (Дані)*** або ***Distance matrix (Матриця відстаней*** (можна використовувати матрицю коефіцієнтів кореляції)***)***. У списку ***Amalgamation (linkage) rule (Правило ієрархічного об'єднання)*** необхідно вказати правило об'єднання в кластери: ***Single Linkage*** (***Критерій "ближнього сусіда" або простий (одиночний) зв'язок)***, ***Complete Linkage (Критерій "далекого сусіда" або повний зв'язок)***, ***Unweighted pair-group average (Критерій середньої відстані, розрахований за формулою простої середньої арифметичної)***, ***Weighted pair-group average (Критерій середньої відстані, розрахований за формулою зваженої середньої арифметичної)***, ***Unweighted pair-group centroid (Критерій центроїда, розрахований без урахування числа об'єктів (статистичної ваги) поєднуваних груп)***, ***Weighted pair-group centroid (median) (Критерій центроїда, розрахований з урахуванням медіани поєднуваних груп)*** і ***Ward's method (Критерій Уорда)***.

У списку ***Distance measure (Міра відстані)*** можна обрати одну з 6 мір: ***Squared Euclidean distances (Квадрат Евклідової відстані)***, ***Euclidean distances (Евклідова відстань)***, ***City-block (Manhattan) distances (Манхеттенівська відстань (відстань міських кварталів))***, ***Chebychev distance metric (Відстань Чебишева)***, ***Power: SUM(ABS(x-y)**p)**1/r (Ступінь:*** $\left(\sum|x-y|^p\right)^{1/r}$***)***, ***Percent disagreement (Відсоток невідповідності)*** та ***1–Pearson r (1–коефіцієнт кореляції Пірсона)***.

Після натискання ***OK*** з'явиться вікно результатів (рис. 12.6). У верхній частині вікна виведена інформація: число змінних, число спостережень, метод кластеризації, правило ієрархічного об'єднання вибрана міра (відстань між об'єктами).

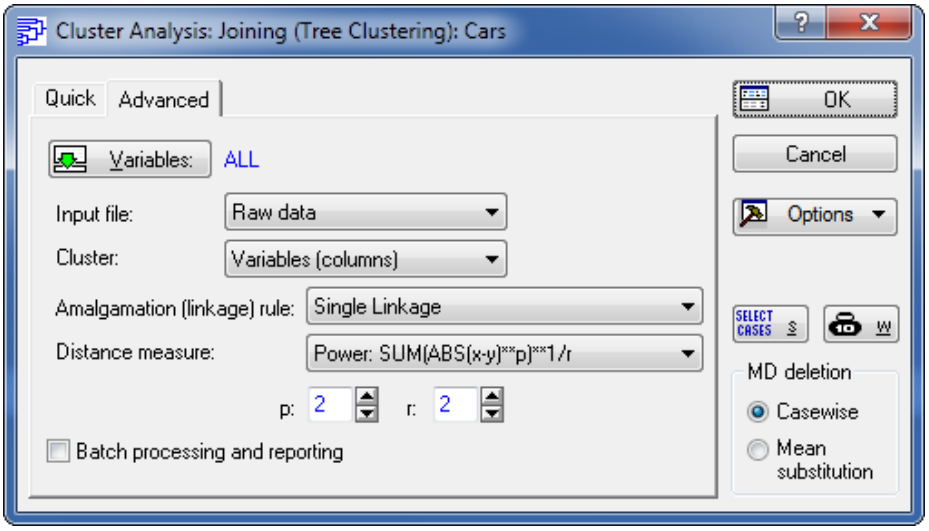


Рис. 12.5. Діалогове вікно кластерного аналізу методом ***Joining tree clustering***

Кнопки в нижній частині вікна на вкладці ***Advanced*** діалогового вікна ***Joining Results*** призначені для аналізу результатів кластеризації:

- ***Horizontal hierarchical tree plot*** – горизонтальна деревовидна діаграма;
- ***Vertical icicle plot*** – вертикальна деревовидна діаграма;
- ***Amalgamation schedule*** – правило об'єднання в кластери;
- ***Graph of amalgamation schedule*** – графік порядку об'єднання;
- ***Distance matrix*** – матриця відстаней;
- ***Descriptive statistics*** – описові статистики.

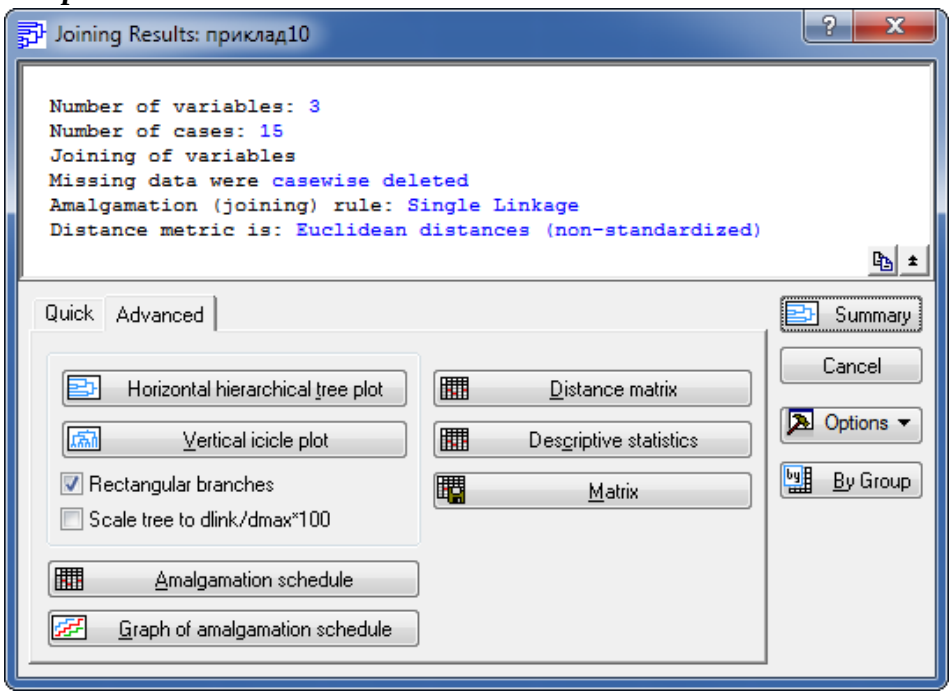


Рис. 12.6 Діалогове вікно ***Joining Results***

Діаграма кластеризації починається з кожного об'єкта в класі (у лівій частині діаграми). Поступово (дуже малими кроками) "слабшає" критерій, що показує, які об'єкти унікальні, а яких немає. Іншими словами, знижується поріг, що відноситься до рішення про об'єднання двох або більше об'єктів у один кластер. У результаті зв'язується все більше і більше число об'єктів і агрегуються (об'єднуються) все більше кластерів, що складаються з елементів, які все сильніше відрізняються. На останньому кроці всі об'єкти остаточно об'єднуються. На

цих діаграмах горизонтальні (чи вертикальні) осі являють собою відстань об'єднання. Так, для кожного вузла у графі (там, де формується новий кластер) можна визначити величину відстані, для якої відповідні елементи зв'язуються в новий єдиний кластер.

Коли дані мають чітку "структуру" в термінах кластерів об'єктів схожих між собою, тоді ця структура може бути відображена в ієрархічному дереві різними гілками. У результаті успішного аналізу методом об'єднання з'являється можливість виявити кластери (гілки) й інтерпретувати їх.

5. Типовий приклад

За результатами вибіркового обстеження умов життя домогосподарств у регіоні (табл. 12.1) виконати групування домогосподарств у однорідні групи.

Таблиця 12.1

№ домогосподарства	*Кількість дітей до 15 років*	*Грошовий місячний дохід, грн.*	*Середньодушові грошові витрати у місяць, грн.*
1	*3*	*1675*	*258*
2	*2*	*2446*	*540*
3	*2*	*2172*	*456*
4	*1*	*2517*	*561*
5	*3*	*1390*	*249*
6	*2*	*1464*	*335*
7	*0*	*1526*	*496*
8	*3*	*1485*	*262*
9	*2*	*1950*	*377*
10	*1*	*1496*	*374*
11	*2*	*2475*	*441*
12	*1*	*1076*	*292*
13	*3*	*1735*	*307*
14	*4*	*1625*	*230*
15	*0*	*1654*	*641*

Розв'язування. *Для здійснення групування в однорідні групи скористаємося кластерним аналізом. Для цього відкриємо модуль* ***Cluster Analisis*** *та виберемо метод –* ***k-means clustering.*** *У вікні, що з'явилося після натискання* ***OK,*** *виберемо всі змінні. Оскільки, ознаки змінюються за рядками, то необхідно вказати у полі* ***Cluster – Cases (rows).*** *Також задамо інші параметри кластеризації (рис. 12.7).*

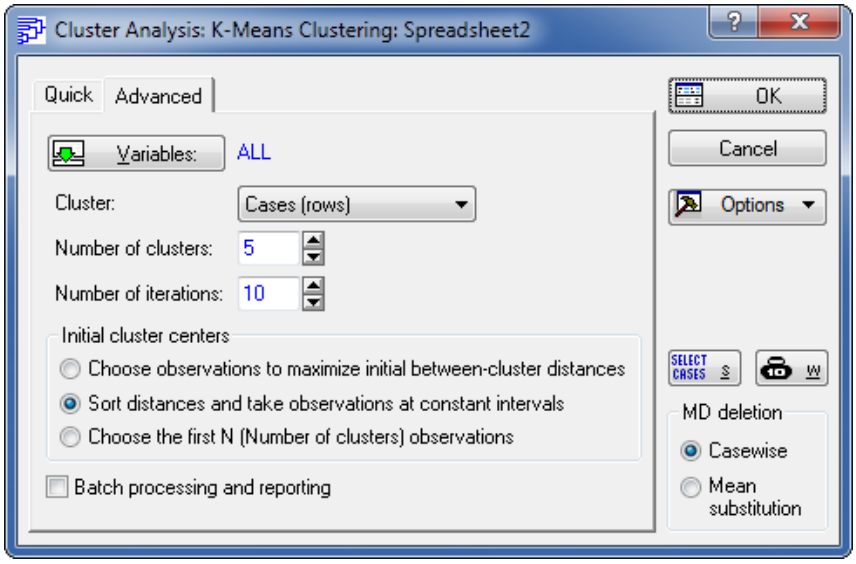


Рис. 12.7. Задання параметрів кластеризації

Натиснемо ОК і перейдемо до вікна результатів кластеризації (рис. 12.8). Щоб визначити, які домогосподарства відносять до побудованих кластерів, натиснемо ***Members of each clusters & distances.***

У результаті буде побудовано 5 таблиць із зазначенням об'єктів кластерів і відстані між центрами кластерів (рис. 12.9).

Отримані 5 кластерів складаються з таких одиниць: перший –з 7 та 15 домогосподарств, другий – з 1, 13, 14 домогосподарств, третій – з 5, 6, 8, 10, 12 домогосподарств, четвертий – з 3 і 9 домогосподарств та п'ятий – з 2, 4, 11 домогосподарств.

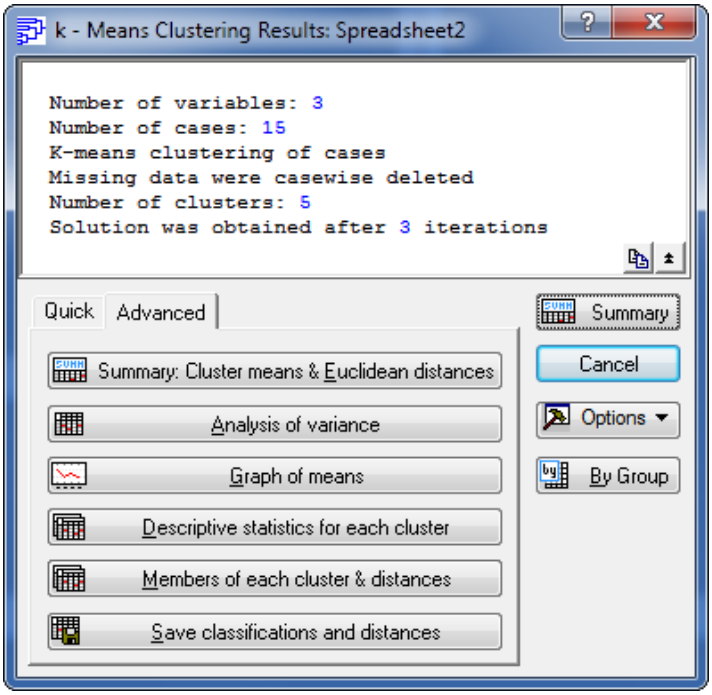


Рис. 12.8. Результати кластерного аналізу методом ***k-means clustering***

	Members of Cluster Number 1 (and Distances from Respective Cluster contains 2 cases
	Distance
7	55,83383
15	55,83383

	Members of Cluster Number 2 (and Distances from Respective Cluster contains 3 cases
	Distance
1	4,48041
13	40,72355
14	36,83245

	Members of Cluster Number 3 and Distances from Respective Cluster contains 5 cases
	Distance
5	31,1630
6	50,8396
8	63,7730
10	77,6273
12	176,8875

	Members of Cluster Number 4 (and Distances from Respective Cluster contains 2 cases
	Distance
3	68,02267
9	68,02267

	Members of Cluster Number 5 and Distances from Respective Cluster contains 3 cases
	Distance
2	24,40780
4	34,77654
11	42,22120

Рис. 12.9. Поділ на кластери

Графічно кластеризація подана на ієрархічному дереві, де по осі абсцис відкладається відстань до центру групування (центру кластеру) та номери об'єктів – по осі ординат. Дерево кластеризації (рис. 12.10) побудуємо за допомогою методу ***Joining tree clustering*** *(щоб повернутися у стартове вікно модуля, необхідно двічі натиснути кнопку* ***Cancel (Скасувати)****). Далі виберемо* ***Horizontal hierarchical tree plot.***

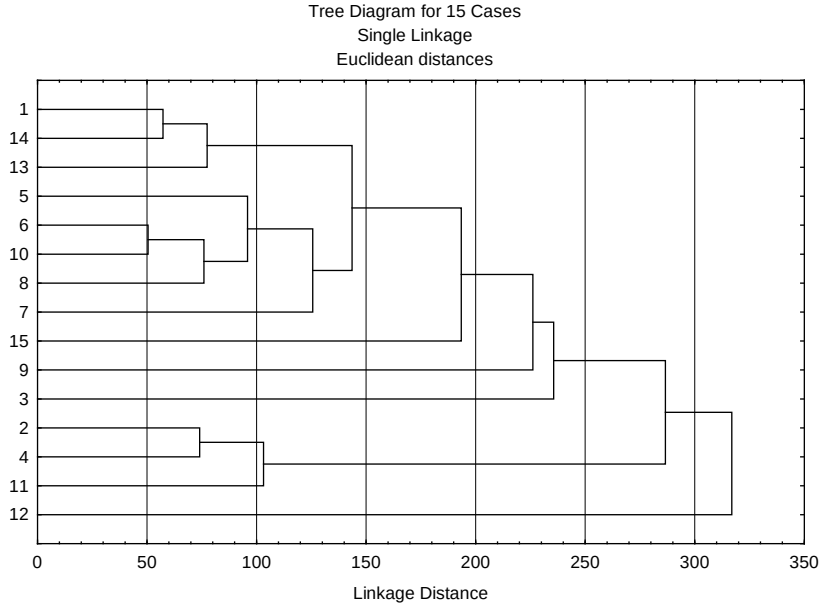


Рис. 12.10. Дерево кластеризації

Завдання для самостійної роботи

12.1. На підприємстві є 16 відділів, зайнятих випуском різноманітної продукції, виконанням робіт, послугами. Оскільки види діяльності, кількість працюючих, рентабельність відділів істотно відрізняться між собою, вирішено згрупувати відділи в кілька однорідних груп, а потім для кожної групи розробити свою систему преміювання. Після ретельного аналізу вибрали чотири ознаки, за допомогою яких описувалися важливі (для зазначеної мети) параметри кожного відділу:

X_1 – вартість основних виробничих фондів, тис. грн.;

X_2 – середньомісячний обсяг робіт відділу, тис. грн.;

X_3 – питома вага робіт/послуг відділу в обсязі всієї виробленої продукції, %;

X_4 – середньомісячний прибуток відділу, тис. грн.

Вихідні дані по відділах наведені в таблиці 12.2.

Таблиця 12.2

№ відділу	*Значення ознак*			
	x_1	x_2	x_3	x_4
1	699	190	53	11
2	532	211	19	42
3	650	152	46	14
4	768	216	67	17
5	67	106	0	32
6	322	397	26	52
7	736	180	49	18
8	501	239	11	60
9	293	391	16	66
10	300	396	29	87
11	73	160	0	22
12	862	199	51	22
13	112	136	0	29
14	289	388	31	74
15	512	195	6	58
16	490	201	9	65

Здійснити кластеризацію відділів за допомогою ієрархічних алгоритмів:

а) використовуючи вихідні дані;

б) використовуючи стандартизовані дані, тобто дані, перетворені за формулою $\frac{x_{ij} - \overline{x}_j}{\sigma_j}$.

Процедуру стандартизації даних можна виконати безпосередньо в таблиці, використовуючи таку послідовність дій: курсор на імені змінної→натиснути праву кнопку миші→у контекстному меню вибрати ***File/Standardize Block (Стандартизувати файл/блок)→Standardize Columns (Стандартизувати стовпець)→OK***.

Здійснити результати кластеризації. За результатами кластеризації знайти число кластерів та їх склад, статистичні характеристики кожного кластера.

Провести кластеризацію, використовуючи метод k-середніх (число кластерів задати 4). Порівняти результати (склад кластерів).

12.2. У таблиці 12.3 наведені значення основних факторів сільськогосподарського виробництва для 20 районів: x_1 – число тракторів на 100 га; x_2 – число зернозбиральних комбайнів на 100 га; x_3 – число знарядь поверхневого обробітку ґрунту на 100 га; x_4 – кількість добрив, що витрачаються на гектар (т/га); x_5 – кількість хімічних засобів захисту рослин, що витрачаються на гектар (ц/га).

Таблиця 12.3

Райони	*Фактори*				
	x_1	x_2	x_3	x_4	x_5
1	1,59	0,26	2,05	0,32	0,14
2	0,34	0,28	0,46	0,59	0,66
3	2,53	0,31	2,46	0,30	0,31
4	4,63	0,40	6,44	0,43	0,59
5	2,16	0,26	2,16	0,39	0,16
6	2,16	0,30	2,69	0,32	0,17
7	0,68	0,29	0,73	0,42	0,23
8	0,35	0,26	0,42	0,21	0,08
9	0,52	0,24	0,49	0,20	0,08
10	3,42	0,31	3,02	1,37	0,73
11	1,78	0,30	3,19	0,73	0,17
12	2,40	0,32	3,30	0,25	0,14
13	9,36	0,40	11,51	0,39	0,38
14	1,72	0,28	2,26	0,82	0,17
15	0,59	0,29	0,60	0,13	0,35
16	0,28	0,26	0,30	0,09	0,15
17	1,64	0,29	1,44	0,20	0,08
18	0,09	0,22	0,05	0,43	0,20
19	0,08	0,25	0,03	0,73	0,20
20	1,36	0,26	0,17	0,99	0,42

Здійснити кластеризацію районів, використовуючи ієрархічний алгоритм (Joining) на основі вихідних і стандартизованих даних, а також кластеризацію за допомогою методу k-середніх (число кластерів задати 3). Порівняти склад кластерів за кожним методом та їх характеристики. Сформулювати висновки.

12.3. За даними таблиці 12.4 класифікувати: 1) об’єкти ієрархічним методом (деревоподібна кластеризація); 2) методом k-середніх (число кластерів – 3). Порівняти склад кластерів та їх характеристики.

Таблиця 12.4

№ з/п	*Країна*	*Число лікарів на 10000 населення*	*Смертність на 10000 населення*	*ВВП за паритетом купівельної спроможності, у % до США*	*Витрати на охорону здоров’я, у % до США*
1	Росія	44.5	84.98	20.4	3.2
2	Австралія	32.5	30.58	71.4	8.5
3	Австрія	33.9	38.42	78.7	9.2
4	Азербайджан	38.8	60.34	12.1	3.3
5	Вірменія	34.4	60.22	10.9	3.2
6	Білорусь	43.6	60.79	20.4	5.4
7	Бельгія	41	29.82	79.7	8.3
8	Болгарія	36.4	70.57	17.3	5.4

№ з/п	*Країна*	*Число лікарів на 10000 населення*	*Смертність на 10000 населення*	*ВВП за паритетом купівельної спроможності, у % до США*	*Витрати на охорону здоров'я, у % до США*
9	Великобританія	17.9	34.51	69.7	7.1
10	Україна	32.1	64.73	24.5	6
11	Німеччина	38.1	36.63	76.2	8.6
12	Греція	41.5	32.84	44.4	5.7
13	Грузія	55	62.64	11.3	3.5
14	Данія	36.7	34.07	79.2	6.7
15	Ірландія	15.8	39.27	57	6.7
16	Іспанія	40.9	28.46	54.8	7.3
17	Італія	49.4	30.27	72.1	8.5
18	Казахстан	38.1	69.04	13.4	3.3
19	Канада	27.6	25.42	79.9	10.2
20	Киргизія	33.2	53.13	11.2	3.4

12.4. За даними таблиці 12.5 класифікувати об'єкти: 1) ієрархічним методом (деревоподібна кластеризація); 2) методом k-середніх (число кластерів – 4). Порівняти склад кластерів та їх характеристики.

Таблиця 12.5

№ з/п	*Країна*	*М'ясо, кг*	*Масло тваринне, кг*	*Цукор, кг*	*Алкоголь, л*	*Фрукти, кг*	*Хлібо-продукти, кг*
1	Росія	55	3,9	30	5	28	124
2	Австралія	100	2,6	47	8,2	121	87
3	Австрія	93	5,3	37	12	146	74
4	Азербайджан	20	4,1	12,4	7,9	52	141
5	Вірменія	20	3,7	4,3	6,5	72	134
6	Білорусь	72	3,6	28	5,4	38	120
7	Бельгія	85	6,9	48	11	83	72
8	Болгарія	65	3	18	9,5	92	156
9	Великобританія	67	3,5	39	8,8	91	91
10	Україна	73	1,7	40	10,9	73	106
11	Німеччина	88	6,8	35	8,1	138	73
12	Греція	83	1	24	8,8	99	108
13	Грузія	21	3,8	36	9,8	55	140
14	Данія	98	5	38	10,3	89	77
15	Ірландія	99	3,3	31	9,6	87	102
16	Іспанія	89	0,4	26	8,95	103	72
17	Італія	84	2,2	27	9,6	169	118
18	Казахстан	61	4,2	19,2	7,2	10	191
19	Канада	98	3,1	44	7,4	123	77
20	Киргизія	46	4,1	23,5	6,7	20	134

12.5. У таблиці 12.6 наведено дані про індекс людського розвитку, розрахований для країн в 2011 році, та основні соціально-економічні показники.

Таблиця 12.6

Країни	*Індекс людського розвитку ООН 2011*	*ВВП на душу населення, дол. США*	*Середня заробітна плата, дол. США*	*Середня пенсія, дол. США*	*Тривалість життя, роки*	*Рівень безробіття, %*
Естонія	0.835	16417	843	305	75	10,2
Литва	0.810	13353	800	290	72	13,7
Латвія	0.805	12725	600	400	72	16,1
Білорусь	0.756	5820	329	194,9	70	0,6
Росія	0.755	13089	774	249,3	68	1,5
Казахстан	0.745	11243	578	152,7	67	0,7
Грузія	0.733	3202	300	40	73	9,3
Україна	0.729	3615	323	150	70	2,2
Вірменія	0.716	3305	295	74,6	74	4,4
Азербайджан	0.700	6915	454	141,5	71	0,8
Туркменістан	0.686	4722	300	70	65	...
Молдова	0.649	1966	280	66,7	69	2,5
Узбекистан	0.641	1380	240	60	68	0,2
Киргистан	0.615	1074	197	49,9	69	2,6
Таджикистан	0.607	934	92	28,4	67	2,6

Здійснити кластеризацію районів, використовуючи ієрархічний алгоритм (Joining) на основі вихідних і стандартизованих даних, а також методом k-середніх (число кластерів – 5). Порівняти склад кластерів та їх характеристики. Сформулювати відповідні висновки.

Список рекомендованої літератури

1. Григорків В.С., Вінничук О.Ю., Григорків М.В., Маханець Л.Л. Статистика: основи теорії та практикум: Навчальний посібник / Григорків В.С., Вінничук О.Ю., Григорків М.В., Маханець Л.Л. Чернівці : Чернівец. нац. ун-т, 2022. 304 с.

2. Фетісов В. С. Пакет статистичного аналізу даних STATISTICA : навч. посіб. Ніжин: НДУ ім. М. Гоголя, 2018. 114 с.

3. Спеціальні розділи математики. Статистичний аналіз даних у середовищі STATISTICA [Електронний ресурс]: навч. посіб. / уклад.: І.М. Джигирей, Д.М. Складанний. Київ : КПІ ім. Ігоря Сікорського, 2019. 74 с.

URL: https://ela.kpi.ua/bitstream/123456789/28228/1/2019srm2.pdf.

4. Лабораторний практикум з навчальної дисципліни "Статистичне моделювання та прогнозування" для студентів напряму підготовки 6.030506 "Прикладна статистика" денної форми навчання / укл. О. В. Раєвнєва, К. А. Стрижиченко, І. В. Чанкіна та ін. – Харків: Вид. ХНЕУ, 2013. 60 с.

5. TIBCO Statistica® User's Guide.

URL: https://docs.tibco.com/pub/stat/14.0.0/doc/html/UsersGuide/GUID-A168AF7A-BC67-4DDF-8CBE-1EE7FE60282D.html.