МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТАВРІЙСЬКИЙ ДЕРЖАВНИЙ АГРОТЕХНОЛОГІЧНИЙ
УНІВЕРСИТЕТ ІМЕНІ ДМИТРА МОТОРНОГО

Віра Малкіна, Ольга Зінов'єва

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Лабораторний практикум

Частина І

Мелітополь

2021

УДК 004.4:004.05

Дозвіл до впровадження та видання надано Вченою радою факультету енергетики і комп'ютерних технологій Таврійського державного агротехнологічного університету імені Дмитра Моторного (протокол № 9 від «26» травня 2021 р.)

Рецензенти:

Мацулевич О.Є. – кандидат технічних наук, доцент кафедри технічної механіки та комп'ютерних технологій проектування ім. В.М. Найдиша, Таврійський державний агротехнологічний університет імені Дмитра Моторного

Строкань О.В. – кандидат технічних наук, доцент кафедри комп'ютерних наук, Таврійський державний агротехнологічний університет імені Дмитра Моторного

Малкіна В. М., Зінов'єва О.Г. Інтелектуальний аналіз даних: Лабораторний практикум. Частина І – Мелітополь: Люкс, 2021. – 150 с.

Лабораторний практикум підготовлено відповідно до програми навчальної дисципліни «Інтелектуальний аналіз даних», яка включена у навчальний план підготовки бакалаврів спеціальності 122 «Комп'ютерні науки» Частина 1 практикуму орієнтована на Змістовий модуль 1 і включає задачі статистичного аналізу даних: первинна обробка статистичних даних, задачі регресійного аналізу та аналізу часових рядів.

УДК 004.4:004.05

ЗМІСТ

ВСТУП

Обробка даних, в тому числі і результатів експерименту, є найважливішим засобом отримання нових знань не тільки в галузі природничих та технічних наук, а й в економіці, соціології, політиці, психології, літературознавстві і в інших галузях. Ці дослідження дають критерії оцінки обґрунтованості та прийнятності на практиці будь-яких теорій і теоретичних припущень. Обробка даних спрямована, як правило, на побудову математичної моделі досліджуваного об'єкта або явища, а також на отримання відповіді на питання: «Чи достовірні наявні дані в межах необхідної точності або допусків?». Сама ж математична модель в залежності від цілей (дослідження, управління, контроль) може бути використана для різних цілей: для предметно-смислового аналізу об'єкта чи явища, прогнозування їх стану в різних умовах функціонування, управління ними в конкретних ситуаціях, оптимізації окремих параметрів, а також для вирішення якихось інших специфічних завдань. Кінцевою метою будь-якої обробки даних є висунення гіпотез про клас і структурі математичної моделі досліджуваного явища, визначення складу і обсягу додаткових вимірів, вибір можливих методів подальшої статистичної обробки та аналіз виконання основних передумов, що лежать в їх основі.

Мета дисципліни “Інтелектуальний аналіз даних” (ІАД) - вивчення методів сучасної обробки даних – інтелектуального аналізу даних (Data Mining), пошуку у необроблених масивах даних раніше невідомих, практично корисних знань та закономірностей, необхідних для прийняття рішень; огляд методів, програмних продуктів та різних інструментальних засобів, що використовуються Data Mining; розгляд практичних прикладів застосування Data Mining; підготовка студентів до самостійної роботи з розв'язання різних економічних задач засобами Data Mining та розробки інтелектуальних систем. Розглядаються такі загальні поняття: статистичні пакети; нейроні мережі; еволюційні методі і алгоритми пошуку логічних закономірностей.

Предметом вивчення навчальної дисципліни є способи побудови математичних моделей виробничих і бізнес-процесів та їх використання для оптимізації зазначених процесів, що знайомить студента з основними проблемами, принципами, правилами, методами, підходами, специфікою та засобами, які використовуються під час статистичної та інтелектуальної обробки даних, отриманих з різноманітних систем

Викладений матеріал у лабораторному практикумі структурований відповідно до навчального плану зі спеціальності 122 «Комп'ютерні науки» для здобувачів ступеня вищої освіти «Бакалавр».

Частина 1 лабораторного практикуму орієнтована на перший змістовий модуль дисципліни і складається з п'яти лабораторних робіт. Даний практикум містить основні відомості та рекомендації, які будуть корисні студентам усіх форм навчання і спрямовані на закріплення студентами отриманих ними на лекційних заняттях теоретичних знань з методів статистичної обробки даних.

Дані методичні вказівки містять основні відомості та рекомендації, які будуть корисні студентам усіх форм навчання і спрямовані на закріплення студентами отриманих ними на лекційних заняттях теоретичних знань з методів інтелектуального аналізу даних. Інструментальною системою для виконання лабораторних робіт є програмний пакет STATISTICA.

МЕТОДИЧНІ ВКАЗІВКИ ДО ВИКОНАННЯ ЛАБОРАТОРНИХ РОБІТ

ЛАБОРАТОРНА РОБОТА №1

ПЕРВИННА ОБРОБКА СТАТИСТИЧНИХ ДАНИХ. ОПИСОВА СТАТИСТИКА

Мета: Навчитися проводити основні поняття математичної статистики у платформі Statistica

1.1 Основні теоретичні відомості

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i}{n} \text{ – середнє вибіркове,} \tag{1.1}$$

n – обсяг вибірки;

$$\overline{D} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n} \text{ – дисперсія;} \tag{1.2}$$

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1} \text{ – виправлена дисперсія;} \tag{1.3}$$

$$\overline{\sigma} = \sqrt{\overline{D}} \text{ – середнє квадратичне відхилення;} \tag{1.5}$$

$$s = \sqrt{S^2} \text{ – виправлене середнє квадратичне відхилення;} \tag{1.6}$$

$$Me = \begin{cases} x_k, & n = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, & n = 2k \end{cases}$$ - медіана (середнє (1.7)

значення ранжируваного ряду за кількістю варіант)

Довірчий інтервал для математичного очікування *a* з надійністю γ:

$$\bar{x} - t_\gamma \frac{s}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{s}{\sqrt{n}} \quad (1.8)$$

де t_γ знаходиться за заданим значенням *n* з надійністю γ (Додаток 1).

$$A_s = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^3}{n\sigma^3}$$ – коефіцієнт асиметрії; (1/9)

$$A_s^* = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{k} (x_i - \bar{x})^3}{s^3}$$ – оцінка коефіцієнта (1/10)

асиметрії;

$$E = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n\sigma^4} - 3$$ – ексцес; (1.11)

$$E^* = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$ – (1.12)

оцінка ексцесу;

$$V = \frac{\bar{\sigma}}{\bar{x}} \cdot 100\%$$ – коефіцієнт варіації. (1.13)

1.2 Завдання для самопідготовки

В процесі підготовки до заняття студент повинен виконати наступні завдання:

а) За допомогою конспекту лекцій і рекомендованої літератури розглянути сутність таких питань:

1) побудова ранжируваних дискретних, динамічних та інтервальних рядів і відповідних їм графіків;

2) обчислення точкових оцінок, що характеризують дискретні, динамічні та інтервальні ряди

1.3 Програма роботи

1) Вивчити теоретичні відомості
2) Проробити контрольний приклад
3) Виконати індивідуальні завдання
4) Оформити звіт. Вимоги до оформлення звіту наведені у п. 1.5
5) Захистити лабораторну роботу. Питання для самоконтролю наведені у п. 1.6

1.4 Порядок виконання роботи

1.4.1 Контрольний приклад

Задача. Отримано дані про врожайність зернових культур(ц/га) у 30 господарствах області:

25,3; 34,2; 38,2; 26,7; 34,3; 38,9; 28,5; 34,4; 39,9; 29,1; 34,7; 40,5; 30,5; 35,1; 41,5; 31,7; 35,8; 42,3; 32,4; 36,7; 44,1; 32,8; 36,8; 44,9; 32,9;37,3; 33,1; 37,6; 33,5; 37,9.

Провести первинну статистичну обробку даних:

а) побудувати інтервальний статистичний ряд;

б) побудувати гістограму;

в) визначити числові характеристики вибірки для згрупованих та незгрупованих даних: $\bar{x}$, $\overline{D}$, $\bar{\sigma}$, A_s, E, $\bar{V}$, As^*, E^*.

Розв’язання

1) Будуємо ранжируваний ряд (табл. 1.1)

Таблиця 1.1 – Ранжируваний ряд

25,3	26,7	28,5	29,1	30,5	31,7
32,4	32,8	32,9	33,1	33,5	34,2
34,3	34,4	34,7	35,1	35,8	36,7
36,8	37,3	37,6	37,9	38,2	38,9
39,9	40,5	41,5	42,3	44,1	44,9

2) Будуємо інтервальний статистичний ряд (виконуємо групування даних):

- знаходимо максимальний і мінімальний елемент вибірки:

$$x_{\max} = 44{,}9;\ x_{\min} = 25{,}3.$$

- визначаємо кількість інтервалів розбиття вибірки:

$$k = 1 + 3{,}2 \cdot \lg n = 1 + 3{,}2 \cdot \lg 30 \approx 5.$$

- за формулою $h = \dfrac{x_{\max} - x_{\min}}{k}$ знаходимо довжину інтервалу розбивки.

У нашому прикладі $n = 30$, $k = 5$, тому

$$h = \frac{44{,}9 - 25{,}3}{5} = 3{,}92;$$

- визначаємо частоти n_i;

- визначаємо відносні частоти P_i^*:

$$P_i^* = \frac{n_i}{n}, \quad \sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} P_i^* = 1$$

- будуємо інтервальний статистичний ряд (таблиця 1.2).

Таблиця 1.2 – Інтервальний статистичний ряд

$x_i - x_{i+1}$	25,3 - 29,22	29,22-33,14	33,14-37,06	37,06-40,98	40,98-44,9
n_i	4	6	9	7	4
P_i^*	0,13	0,20	0,30	0,23	0,13
$\frac{P_i^*}{h}$	0,03	0,05	0,08	0,06	0,03

3) Графічним зображенням інтервального ряду є гістограма. Будуємо гістограму відносних частот (див. рисунок 1.1):

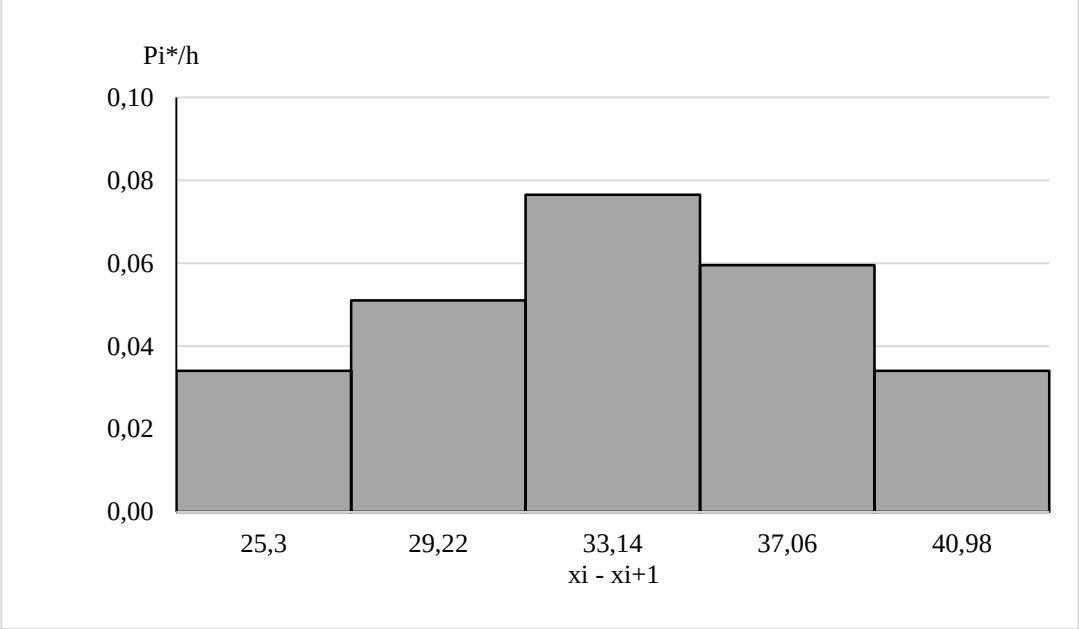

Рисунок 1.1 – Гістограма щільності відносних частот

4) Обчислимо середню врожайність зернових культур для незгрупованих даних за формулою (1.1):

$$\overline{x} = \frac{25{,}3 + 26{,}7 + \ldots + 44{,}9}{30} = 35{,}38;$$

5) Знаходимо медіану за формулою (1.7):

$$Me = \frac{x_{15} + x_{16}}{2} = \frac{34,7 + 35,1}{2} = 34,9;$$

6) Знайдемо дисперсію $\overline{D}$ та виправлену дисперсію S^2 :для незгрупованих даних за формулами (1.2) та (1.3) відповідно:

$$\overline{D} = \frac{25,3^2 + 26,7^2 + 28,5^2 + ... + 44,9^2}{30} - 35,38^2 = 22,44$$

$$S^2 = \frac{(25,3 - 35,38)^2 + (26,7 - 35,38)^2 + ... + (44,9 - 35,38)^2}{29} = 23,29$$

Зауважимо, що виправлену дисперсію, також, можна знайти за формулою

$$S^2 = \frac{\overline{D} \cdot n}{n-1}$$

$$S^2 = \frac{22,4 \cdot 30}{30-1} = 23,29$$

7) Обчислюємо середнє квадратичне відхилення (1.5) та виправлене середнє квадратичне відхилення (1.6) для незгрупованих даних:

$$\overline{\sigma} = \sqrt{22,4} = 4,73;$$

$$s = \sqrt{23,2} = 4,81;$$

8) Визначимо коефіцієнт асиметрії для незгрупованих даних за формулою (1.9):

$$As = \frac{(25,3 - 35,38)^3 + (26,7 - 35,38)^3 + ... + (44,9 - 35,38)^3}{30 * 4.73^3} = -0,029$$

9) Визначимо ексцес для незгрупованих даних за формулою (1.11):

$$E = \frac{(25,3 - 35,38)^4 + (26,7 - 35,38)^4 + ... + (44,9 - 35,38)^4}{30 * 4.73^4} - 3 = -0,36$$

10) Визначимо оцінку коефіцієнту асиметрії за формулою (1.10):

$$As^* = \frac{30}{29*28}\frac{(25{,}3-35{,}38)^3+(26{,}7-35{,}38)^3+\ldots+(44{,}9-35{,}38)^3}{4.81^3} = -0.03$$

Так як, $A_s < 0$, спостерігається лівостороння скошеність ряду.

11) Визначимо оцінку ексцесу за формулою (1.12):

$$E^* = \frac{30*31}{29*28*27} * \frac{(25{,}3-35{,}38)^4+(26{,}7-35{,}38)^4+\ldots+(44{,}9-35{,}38)^4}{4.81^4} -$$

$$-3*\frac{29^2}{28*27} = -0.20$$

Так як, $E < 0$, ряд плосковершинний у порівнянні з рядом, розподіленим по нормальному закону.

12) Визначимо коефіцієнт варіації за формулою (1.13):

$$V = \frac{4{,}66}{35{,}38} = 13{,}16\%$$

13) Довірчий інтервал для математичного очікування *a:*

$$35{,}38 - t_\gamma \frac{4.81}{\sqrt{30}} < a < 35{,}38 + t_\gamma \frac{4.81}{\sqrt{30}},$$

де $t_\gamma = 2{,}045$ знаходиться у відповідності з Додатком 1 за заданими $n = 30$ та $\gamma = 0{,}95$.

Таким чином, $33{,}58 < a < 37{,}18$

1.4.2 Розрахунок в середовищі MS Excel

1) Обчислити описові статистики за допомогою процедури Аналіз даних (MS Excel):

- Дані $\Rightarrow$ Аналіз Даних… $\Rightarrow$ Описова статистика $\Rightarrow$ ОК; (Для ранніх версій MS Excel: Сервіс - Аналіз Даних…Описовастатистика)

- вдіалоговому вікні Описова статистика виконати наступне:

- внести у вікно редагування Вхідний інтервал діапазон A2:A31;

- встановити перемикач Групуванняв положення По стовпцям;

- зняти прапорець Мітки в першому рядку;

- встановити перемикач Параметри виводу в положення Новий робочий лист;

- встановити прапорець Ітогова статистика;

- “клік” на кнопці ОК.

2) Проконтролювати свою роботу і відформатувати таблицю в такий спосіб (див. таблицю 1.3):

Таблиця 1.3 – Описова статистика

Описова статистика	
Середнє	35,38667
Стандартна помилка	0,879821
Медіана	34,9
Мода	#Н/Д
Стандартне відхилення	4,818981
Дисперсія вибіркова	23,22257
Ексцес	-0,20259
Асиметричність	-0,03112
Інтервал	19,6
Мінімум	25,3
Максимум	44,9
Сума	1061,6
Рахунок	30

Висновок. Аналіз вибірки дозволяє зробити висновок, що для генеральної сукупності:

– врожайність зернових культур дорівнює, в середньому, $35{,}4 \pm 4{,}81$ ц/га,

– ряд є плосковершинним у порівнянні з рядом, розподіленим за нормальним законом (E<0), тобто значення інтервальних частот менші ніж значення відповідних частот для ряду, розподіленому за нормальним законом

– має лівосторонню скошеність (As<0), що означає – кількість значень вибірки, до середнього 35,4 менше кількості значень після середнього значення 35,4.

– має значне розсіювання ($V = 13{,}39\%$).

– кількість даних менших, ніж медіана $Me = 34{,}9$ і більш, ніж $Me = 34{,}9$ однакова і дорівнює 15.

1.4.3 Первинний статистичний аналіз за допомогою пакету Statistica

1) Вводимо значення змінної згідно варіанту

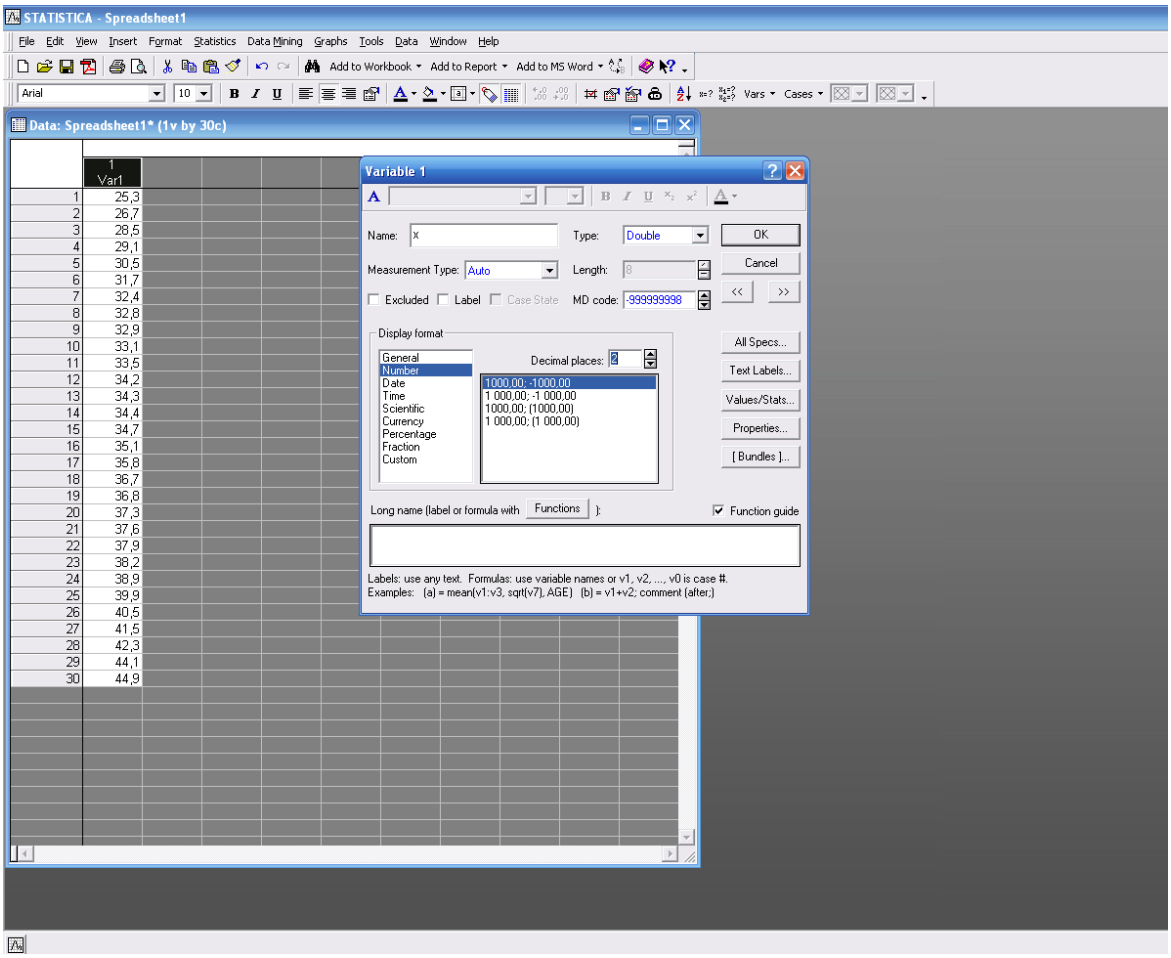

Рисунок 1.2 – Введення даних

2) Для виконання різних видів статистичного аналізу необхідно вибрати командуАналіз (Statistics)(рис. 1.3).

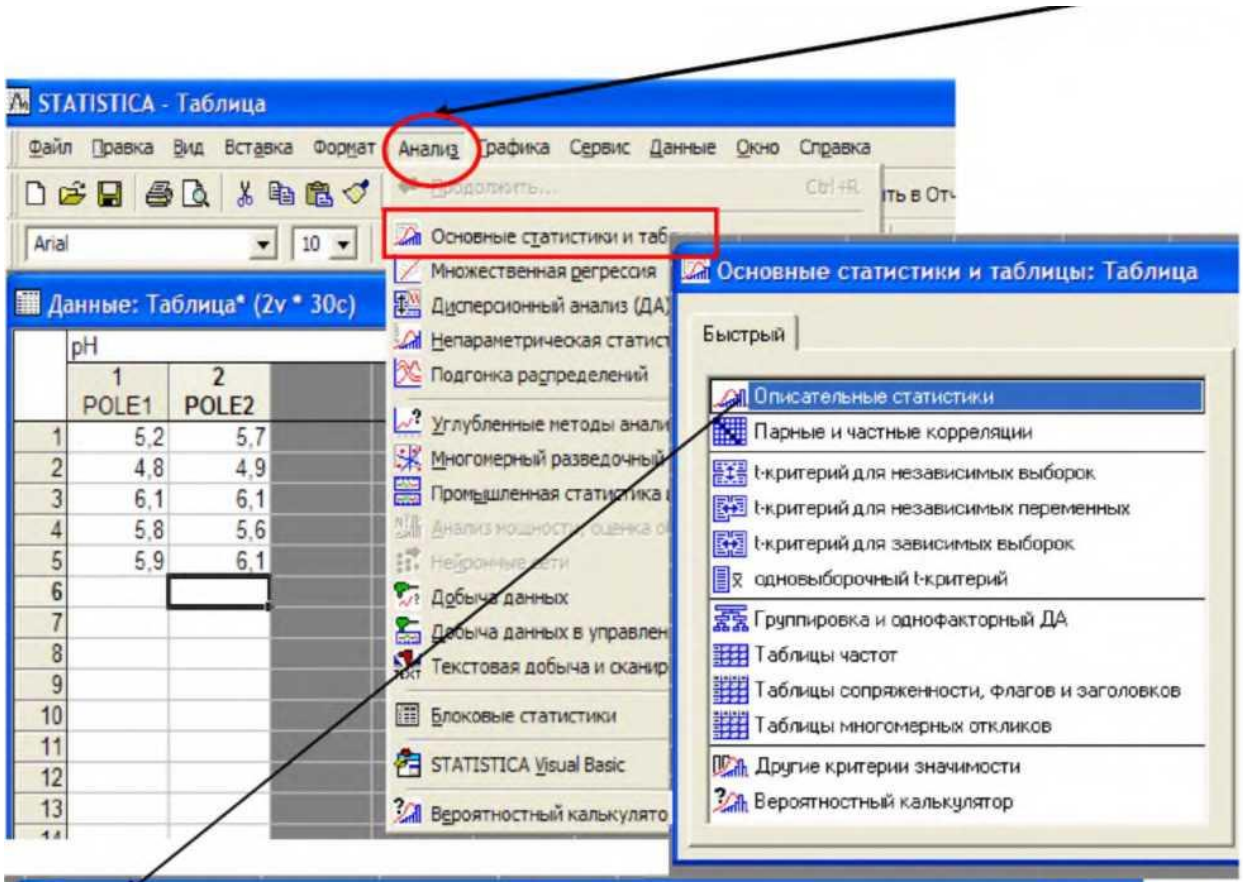

Рисунок1.3 - Виконання статистичного аналізу

3) Клік по кнопці Змінні(Variables) відкриває список змінних, для яких потрібно провести аналіз. Виділення необхідних змінних виконується звичайним для Windows способом. Крім того, можна просто ввести номери змінних у віконці, причому, якщо вони йдуть підряд, то просто вводять номер початкової й кінцевоїзмінної через дефіс. У протилежному випадку номера вводяться через кому або кліками по імені змінної при натиснутій кнопці **Ctrl.**

4) Для завдання необхідних статистик необхідно вибрати опціюДодатково (Advanced)і поставити прапорці у віконцях необхідних показників. Натиснути ОК**.**

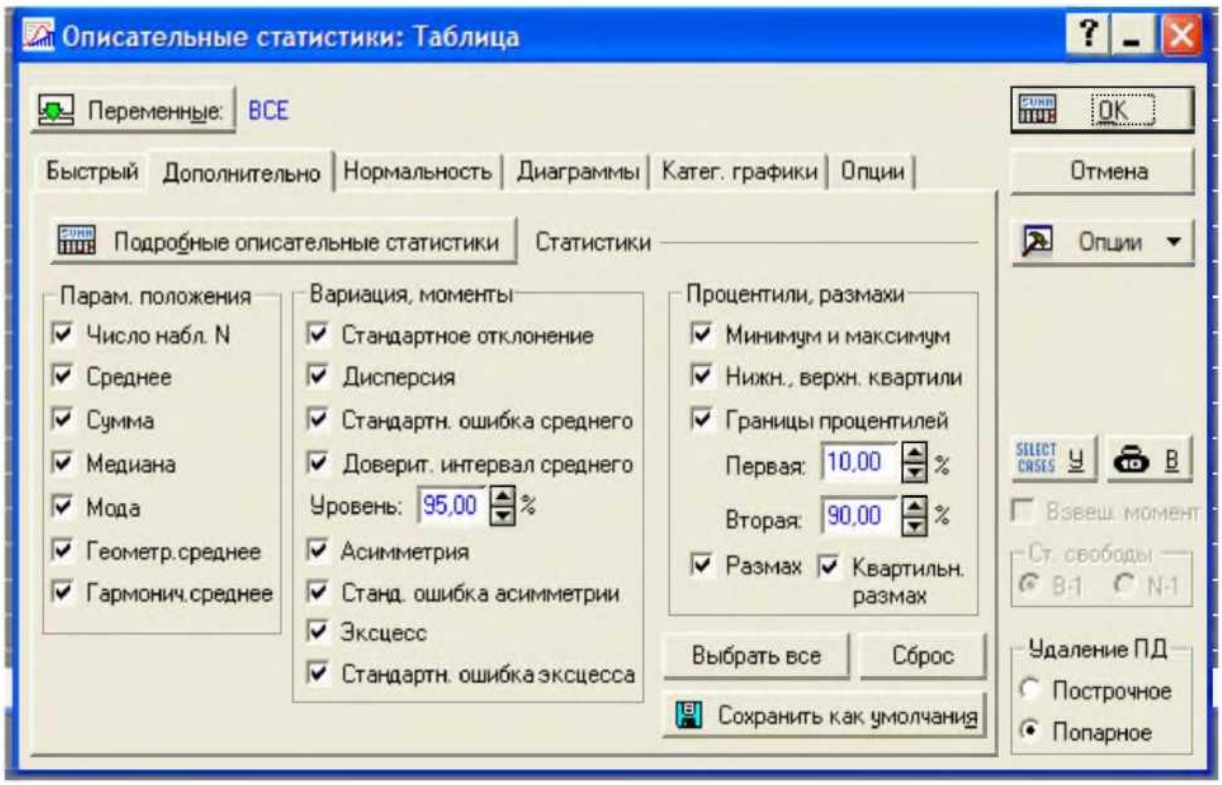

Рисунок 1.4 – Описові статистики

5) Обрати наступні статистики: об'єм вибірки (Valid N), середнє (Mean), медіана (Median), мінімум (Minimum), максимум (Maximum), нижній квартиль(LowerQuartile), верхній квартиль (UpperQuartile), виправлена дисперсія (Variance), виправлене середньє квадратичне відхилення (стандартне відхилення) (Std.Dev.), асиметрія (Skewness), ексцес (Kurtosis), коефіцієнт варіації. Натисніть OK(Summary).

6) Результати обчислень розміщаються в робочу книгу (Workbook)