МІНІСТЕРСТВО ВНУТРІШНІХ СПРАВ УКРАЇНИ

Харківський національний університет внутрішніх справ

Кафедра інформаційних технологій та кібербезпеки, факультет № 4

МЕТОДИЧНІ МАТЕРІАЛИ

ДО ПРАКТИЧНИХ ЗАНЯТЬ

з навчальної дисципліни **«Моделі, методи та засоби аналітичної обробки великих масивів даних»**
вибіркових компонент
освітньої програми *другого*(*магістерського*) рівня вищої освіти
125 «Кібербезпека» («Безпека інформаційних та комунікаційних систем»)

м. Харків 2020

ЗАТВЕРДЖЕНО
Науково-методичною радою
Харківського національного
університету внутрішніх справ
Протокол від 23.09.2020 № 9

СХВАЛЕНО
Вченою радою факультету № 4
Протокол від 16.09.2020 № 5

ПОГОДЖЕНО
Секцією Науково-методичної ради
ХНУВС з технічних дисциплін

Протокол від 18.09.2020 № 5

Розглянуто на засіданні кафедри інформаційних технологій та кібербезпеки (протокол від 15.09.2020 № 16

Розробники:
1. *Професор. кафедри, к.т.н., доцент Струков В.М.*

Рецензенти:
1. *д.т.н., професор Зацеркляний М.М.,*
2. *доцент кафедри програмної інженерії ХНУРЕ, кандидат технічних наук, доцент Лановий О.Ф.*

1. Розподіл часу навчальної дисципліни за темами

Номер та назва навчальної теми	Кількість годин, відведених на вивчення навчальної дисципліни						Вид контролю
	Всього	з них:					
		лекції	Семінарські заняття	Практичні заняття	Лабораторні заняття	Самостійна робота	
Семестр № 1							
Тема № 1. Актуальність і тенденції аналізу Великих Даних у правоохоронній сфері Проблема обробки великих масивів даних.	**18**	**4**	**4**			**10**	**к/р**
Тема № 2: Задачі і етапи опрацювання великих даних.	**12**	**2**	**2**			**8**	
Тема № 3: Елементи теорії множин і теорії графів.	**18**	**4**	**2**			**12**	
Тема № 4.Статистичне дослідження великих даних.	**36**	**4**		**8**		**24**	**к/р**
Тема № 5: Алгоритми ієрархічної кластеризації.	**30**	**4**	**2**	**4**		**20**	
Тема № 6: Методи і алгоритми чіткої кластеризації.	**34**	**4**	**2**	**4**		**24**	
Тема № 7: Інструментальні засоби обробки Великих Даних у правоохоронній діяльності.	**32**	**6**	**4**			**22**	
Всього за семестр № 1:	**180**	**28**	**16**	**16**		**120**	**екзамен**
Всього по дисципліні	**180**	**28**	**16**	**16**		**120**	

2. Методичні вказівки до практичних занять

Тема № 4. Статистичне дослідження великих даних.

Практичне заняття 1. Показники ряду динаміки.

Навчальна мета заняття: сформувати у студентів навички застосування рядів динаміки для аналізу Великих Даних.

Час проведення: 2 год.

Навчальні питання:

1. Абсолютні показники.
2. Відносні показники.
3. Середні показники ряду динаміки.

Література:

Основна.

1. Інформаційні технології у правоохоронній діяльності. Частина 1: Високотехнологічні тренди у правоохоронній сфері зарубіжних країн: навч.

посіб. / Харків. Нац. Ун-т внутр. Справ; [В.М. Струков, Д.В. Узлов, Ю.В. Гнусов та ін.] ; за заг. ред. канд. техн. наук, доц. В.М. Струкова. Харків : ТОВ «ДІСА ПЛЮС», 2020. 276 с.

2. Кубрак В. П. Правова статистика: навч. посіб. / В. П. Кубрак. – МВС України, Харк. нац. ун-т внутр.справ. – Харків: 2017. – 194 с.

3. Конспект лекцій.

Додаткова.

4. Зацеркляний М.М. Інформаційні технології у правозастосовній діяльності: Навч. посібник / М.М. Зацеркляний, В.М. Струков. : Х.: ТОВ „Східно-регіональний центр гуманітарно-освітніх ініціатив”; 2010. 332 с.

5. Зацеркляний М.М. Основи комп’ютерних технологій для економістів: Навч. посібник / М.М. Зацеркляний, О.Ф. Мельников, В.М. Струков. – К.: ВД „Професіонал”, 2006 р. – 672 с.

Інформаційні ресурси в Інтернеті.

6. Gartner: Топ-10 стратегічних трендів розвитку технологій у 2020 році: сайт. URL: https://ain.ua/2018/10/26/gartner-top-10-trendov-razvitiya-texnologij/ (дата звернення: 25.10.2019).

Завдання: Відповідно з даними виданого Вам варіанта розрахуйте абсолютні, відносні та середні показники ряду динаміки, побудуйте стовпчикову діаграму кількості зареєстрованих грабежів, зробіть висновок щодо тенденції розвитку грабежів.

План проведення заняття.

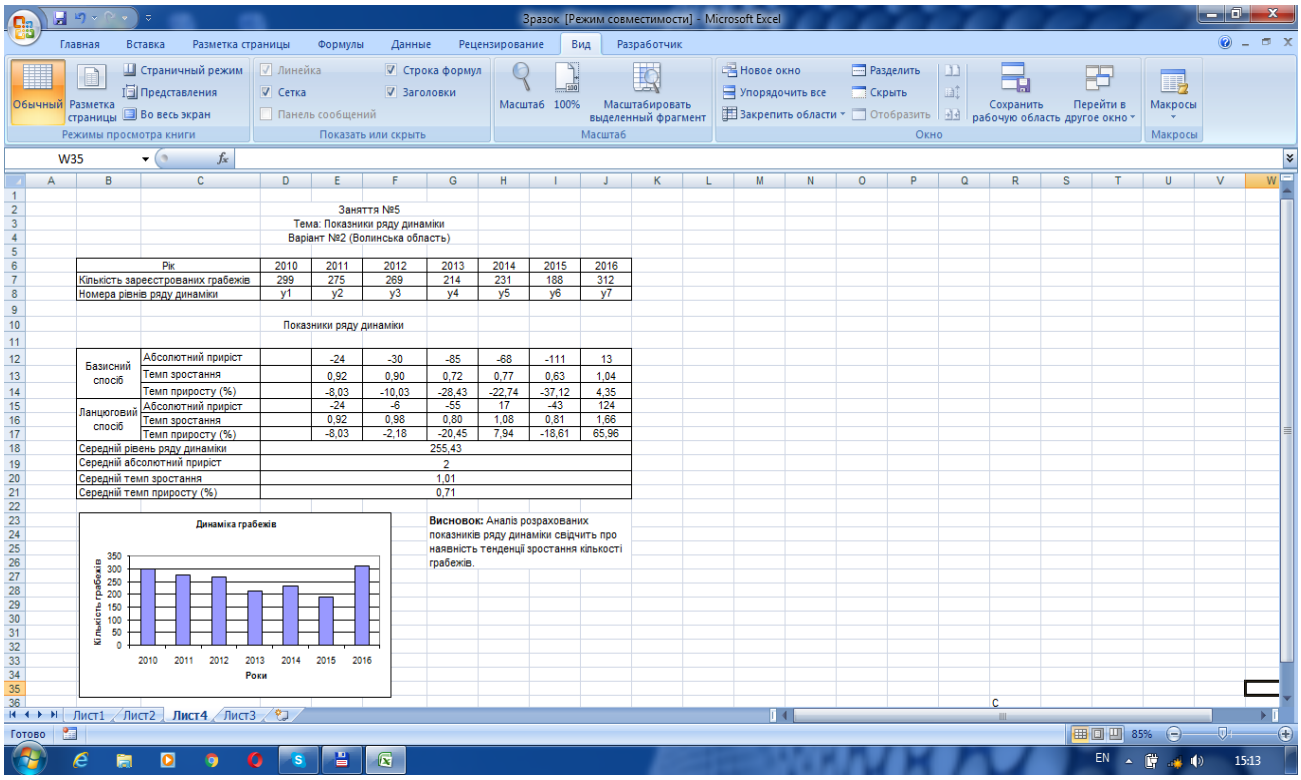
1. Запустіть програму **Excel**.
2. Побудуйте ряд динаміки, вказаний в Вашому варіанті.
3. Розрахуйте показники ряду динаміки базисним способом.

3.1. В комірках **E12:J12** розрахуйте *абсолютні прирости* за кожний рік за формулою: $A^{б}_{i}=y_i - y_б$,

де y_i – кількість грабежів за кожний рік (з 2011 по 2016), а $y_б$ - кількість грабежів за **2010** рік.

3. 2. В комірках **E13:-J13** розрахуйте *темпи зростання* за кожний рік за формулою: $Тзр^{б}_{i} = \frac{y_i}{y_б}$

3.3. В комірках **E14:J14** розрахуйте *темпи приросту* за кожний рік за формулою: $Тпр^{б}i = \frac{A^{б}i}{y_б} \cdot 100\%$



4. Розрахуйте показники ряду динаміки ланцюговим способом.

4.1. В комірках **E15:J15** розрахуйте *абсолютні прирости* за кожний рік за формулою: $A^{л}_i = y_i - y_{i-1}$

де y_i – кількість грабежів за кожний рік (з 2011 по 2016), а y_{i-1} - кількість грабежів за кожний попередній рік.

4.2. В комірках **E16:J16** розрахуйте *темпи зростання* за кожний рік за формулою: $Тзр^{л}_i = \frac{y_i}{y_{i-1}}$

4.3. В комірках **E17:J17** розрахуйте *темпи приросту* за кожний рік за формулою: $Тпр^{л}i = \frac{A^{л}i}{y_{i-1}} \cdot 100\%$

5. В об'єднаних комірках **D18:J18** розрахуйте *середній рівень* ряду динаміки за формулою: $\bar{y} = \frac{y_1 + y_2 + y_3 + ... + y_n}{n}$

6. В об'єднаних комірках **D19:J19** розрахуйте *середній абсолютний приріст* за формулою: $\bar{A} = \frac{A^{л}_2 + A^{л}_3 + ... + A^{л}_n}{n-1}$

7. В об'єднаних комірках **D20:J20** розрахуйте *середній темп зростання* за формулою: $\bar{T}_{зр} = \sqrt[n-1]{Тзр^{л}_2 \cdot Тзр^{л}_3 \cdot ... \cdot Тзр^{л}_n}$

Для цього наберіть формулу: **=(E16*F16*G16*H16*I16*J16)^(1/6)**.

8. В об'єднаних комірках **D21:J21** розрахуйте *середній темп приросту* за формулою: $\bar{T}_{пр} = (\bar{T}_{зр} - 1) \cdot 100\%$

9. Побудуйте гістограму “**Динаміка грабежів**”, яка характеризує зміну кількості зареєстрованих грабежів за період з 2010 по 2016 роки. Для цього:

9.1. Виділіть комірки **D7:J7** (кількість зареєстрованих грабежів).

9.2. Виберіть у меню вкладку **Вставка**, потім в області **Диаграмма** натисніть піктограму **Гистограмма** і виберіть діаграму **Гистограмма с группировкой**.

9.3. В області **Данные** натисніть кнопку **Выбрать данные.**

9.4. Під написом **Подписи горизонтальной оси (категории)** натисніть кнопку **Изменить**.

9.5. У відкрите поле **Подписи оси** введіть роки шляхом виділення комірок **D6:J6 і** натисніть **Ок.**

9.6. Виберіть у меню вкладку **Макет**, у вікні **Подписи** натисніть команду **Название диаграммы** вкажіть **Динаміка грабежів.**

9.7. Натисніть команду **Название осей.** Натисніть команду **Название основной горизонтальной оси**, після цього – **Название под осью**. В полі **Название оси** вкажіть **Роки**.

9.8. Натисніть команду **Название основной вертикальной оси**, у відкритому вікні вкажіть **Повёрнутое название**.

9.9. У відкритому вікні **Название оси** вкажіть **Кількість грабежів**.

9.10. У вікні **Подписи** натисніть команду **Легенда**. У відкритому вікні вкажіть **Нет** (Не добавлять легенду).

9.11. Отриману діаграму встановіть у необхідне місце.

10. **Зробіть висновок** щодо тенденції розвитку грабежів.

Тема № 4. Статистичне дослідження великих даних.

Практичне заняття 2. Середні показники та показники варіації.

Навчальна мета заняття: сформувати у студентів навички обчислювати середні показники та показники варіації, використовуючи функції Microsoft Excel.

Час проведення: 2 год.

Навчальні питання:

1. Середні показники.
2. Мода та медіана.
3. Показники варіації ознаки.

Література:

Основна.

1. Інформаційні технології у правоохоронній діяльності. Частина 1: Високотехнологічні тренди у правоохоронній сфері зарубіжних країн: навч. посіб. / Харків. Нац. Ун-т внутр. Справ; [В.М. Струков, Д.В. Узлов, Ю.В. Гнусов та ін.] ; за заг. ред. канд. техн. наук, доц. В.М. Струкова. Харків : ТОВ «ДІСА ПЛЮС», 2020. 276 с.

2. Кубрак В. П. Правова статистика: навч. посіб. / В. П. Кубрак. – МВС України, Харк. нац. ун-т внутр.справ. – Харків: 2017. – 194 с.

3. Конспект лекцій.

Додаткова.

4. Зацеркляний М.М. Інформаційні технології у правозастосовній діяльності: Навч. посібник / М.М. Зацеркляний, В.М. Струков. : Х.: ТОВ „Східно-регіональний центр гуманітарно-освітніх ініціатив"; 2010. 332 с.

5. Зацеркляний М.М. Основи комп'ютерних технологій для економістів: Навч. посібник / М.М. Зацеркляний, О.Ф. Мельников, В.М. Струков. – К.: ВД „Професіонал", 2006 р. – 672 с.

Інформаційні ресурси в Інтернеті.

6. Gartner: Топ-10 стратегічних трендів розвитку технологій у 2020 році: сайт. URL: https://ain.ua/2018/10/26/gartner-top-10-trendov-razvitiya-texnologij/ (дата звернення: 25.10.2019).

План проведення заняття.

1. Запустіть програму Microsoft Excel.

2. Розрахуйте показники, вказані в таблиці 1. Для цього:

2.1. **Побудуйте статистичну таблицю 1.** Вік осіб, які вчинили злочини, візьміть із **завдання 1**, вказаного у Вашому варіанті. Для того, щоб вказати ступінь **2** у найменуванні показника **(**$X\text{-}X_{сер}$**)²** наберіть після закриваючої дужки цифру **2** і виділіть її. У вкладці **Главная** в пол**і Шрифт** натисніть лівою клавішею миші **стрілочку,** відкриється вікно **Формат ячеек,** у вікні **Видоизменение** увімкніть перемикач **надстрочный.** .

2.2. В комірці D15 розрахуйте середній вік осіб в групі **(Хсер)**. Для цього знайдіть у вкладці **Формулы** функцію "**СРЗНАЧ**", натисніть її, виділіть комірки B7:B11(віки усіх осіб) і натисніть кнопку **Ок.**

2.3. В комірці D16 розрахуйте **середнє квадратичне відхилення**. Формула середнього квадратичного відхилення

$$\sigma = \sqrt{\frac{(X_1 - X_{сер})^2 + (X_2 - X_{сер})^2 + ... + (X_n - X_{сер})^2}{n}}$$

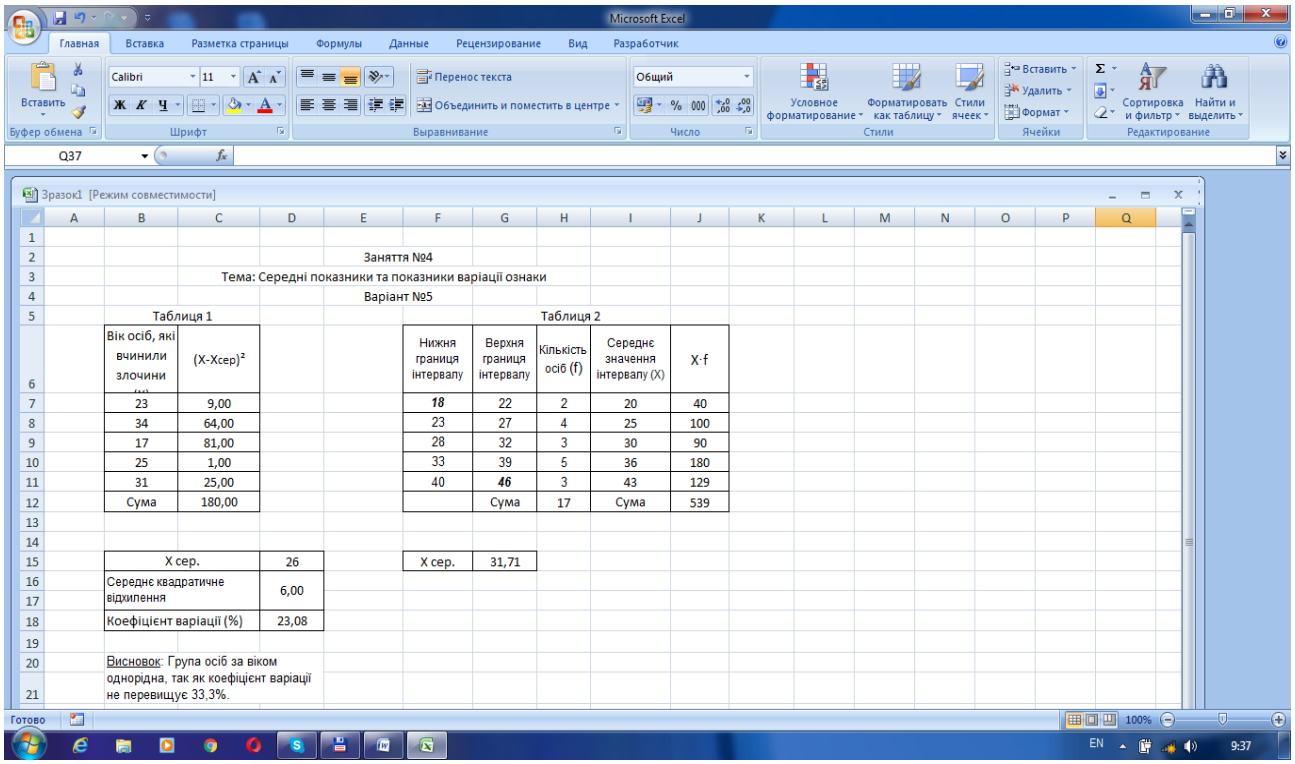
2.4. Для розрахунку <u>середнього</u> квадратичного відхилення спочатку в комірках C7:C11 розрахуйте **квадратичні відхилення** *(* $X_1\text{-}X_{сер})^2$, *(* $X_2\text{-}X_{сер})^2$, *...(* $X_n\text{-}X_{сер})^2$, які знаходяться під коренем, за формулою: **=(*X* - *X*$_{сер}$)^2.**

2.5. В комірці C12 розрахуйте суму квадратичних відхилень.

2.6. Для розрахунку в комірці D16 середнього квадратичного відхилення знайдіть і використайте у вкладці **Формулы** функцію "**КОРЕНЬ**". Формула для розрахунку буде мати вигляд: = **КОРЕНЬ(C12/5).**

2.7. В комірці D18 розрахуйте **коефіцієнт варіації**.

2.8. Зробіть висновок: однорідна за віком група осіб в таблиці 1 чи ні, і **<u>чому.</u>**



Microsoft Excel

Зразок1 [Режим совместимости]

Заняття №4

Тема: Середні показники та показники варіації ознаки

Варіант №5

Таблиця 1

Вік осіб, які вчинили злочини (x)	(X-Xсер)²
23	9,00
34	64,00
17	81,00
25	1,00
31	25,00
Сума	180,00

Таблиця 2

Нижня границя інтервалу	Верхня границя інтервалу	Кількість осіб (f)	Середнє значення інтервалу (X)	X·f
18	22	2	20	40
23	27	4	25	100
28	32	3	30	90
33	39	5	36	180
40	46	3	43	129
	Сума	17	Сума	539

X сер.	26
Середнє квадратичне відхилення	6,00
Коефіцієнт варіації (%)	23,08

X сер.	31,71

Висновок: Група осіб за віком однорідна, так як коефіцієнт варіації не перевищує 33,3%.

3. Розрахуйте показники, вказані в таблиці 2. **Для цього:**

3.1. Побудуйте статистичну таблицю 2. Дані, які вказані в таблиці, візьміть із **завдання 2**, вказаного в Вашому варіанті.

3.2. Визначить нижню границю першого інтервалу і верхню границю п'ятого інтервалу.

3.3. В комірках I7:I11 визначить середнє значення віку (X) в кожному інтервалі.

3.4. В комірці G15 розрахуйте середній вік осіб **(Xсер)** за формулою середньої арифметичної зваженої: $x_{сер} = \frac{\sum(x \cdot f)}{\sum f}$. Для цього спочатку розрахуйте окремо чисельник і знаменник цієї формули.

3.5. Для розрахунку чисельника визначить в комірках J7:J11 добутки **X·f** в кожному інтервалі, а потім в комірці J12 визначить їх суму $\sum(X \cdot f)$.

3.6. Для розрахунку знаменника визначить в комірці H12 загальну кількість осіб в групі $\sum f$.

3.7. Розрахуйте середній вік осіб, які вчинили злочини, за формулою середньої арифметичної зваженої.

Тема № 4. Статистичне дослідження великих даних.

Практичне заняття 3. Визначення тенденції розвитку правопорушень.

Навчальна мета заняття: навчитися визначати тенденцію розвитку правопорушень, використовуючи функції Microsoft Excel.

Час проведення: 2 год.

Навчальні питання:

1. Виявлення основної тенденції зміни в часі досліджуваного явища.

2. Визначення сезонних коливань рівнів ряду динаміки.

3. Показники варіації ознаки.

Література:

Основна.

1. Інформаційні технології у правоохоронній діяльності. Частина 1: Високотехнологічні тренди у правоохоронній сфері зарубіжних країн: навч. посіб. / Харків. Нац. Ун-т внутр. Справ; [В.М. Струков, Д.В. Узлов, Ю.В. Гнусов та ін.] ; за заг. ред. канд. техн. наук, доц. В.М. Струкова. Харків : ТОВ «ДІСА ПЛЮС», 2020. 276 с.

2. Кубрак В. П. Правова статистика: навч. посіб. / В. П. Кубрак. – МВС України, Харк. нац. ун-т внутр.справ. – Харків: 2017. – 194 с.

3. Конспект лекцій.

Додаткова.

4. Зацеркляний М.М. Інформаційні технології у правозастосовній діяльності: Навч. посібник / М.М. Зацеркляний, В.М. Струков. : Х.: ТОВ „Східно-регіональний центр гуманітарно-освітніх ініціатив”; 2010. 332 с.

5. Зацеркляний М.М. Основи комп’ютерних технологій для економістів: Навч. посібник / М.М. Зацеркляний, О.Ф. Мельников, В.М. Струков. – К.: ВД „Професіонал”, 2006 р. – 672 с.

Інформаційні ресурси в Інтернеті.

6. Gartner: Топ-10 стратегічних трендів розвитку технологій у 2020 році: сайт. URL: https://ain.ua/2018/10/26/gartner-top-10-trendov-razvitiya-texnologij/ (дата звернення: 25.10.2019).

Завдання:

1) В *завданні №1* визначить тенденцію зміни кількості зареєстрованих злочинів, використовуючи вирівнювання ряду динаміки методом укрупнення інтервалів часу і методом обчислення ковзної середньої, зробіть відповідний висновок.

2) В *завданні №2* визначить сезонні коливання злочинів, розрахувавши індекси сезонності, зробіть відповідний висновок.

3) Побудувати графік, який характеризує динаміку злочинів в першому завданні.

4) Побудувати графік, який характеризує сезонні коливання злочинів (сезонну хвилю) в другому завданні.

План проведення заняття.

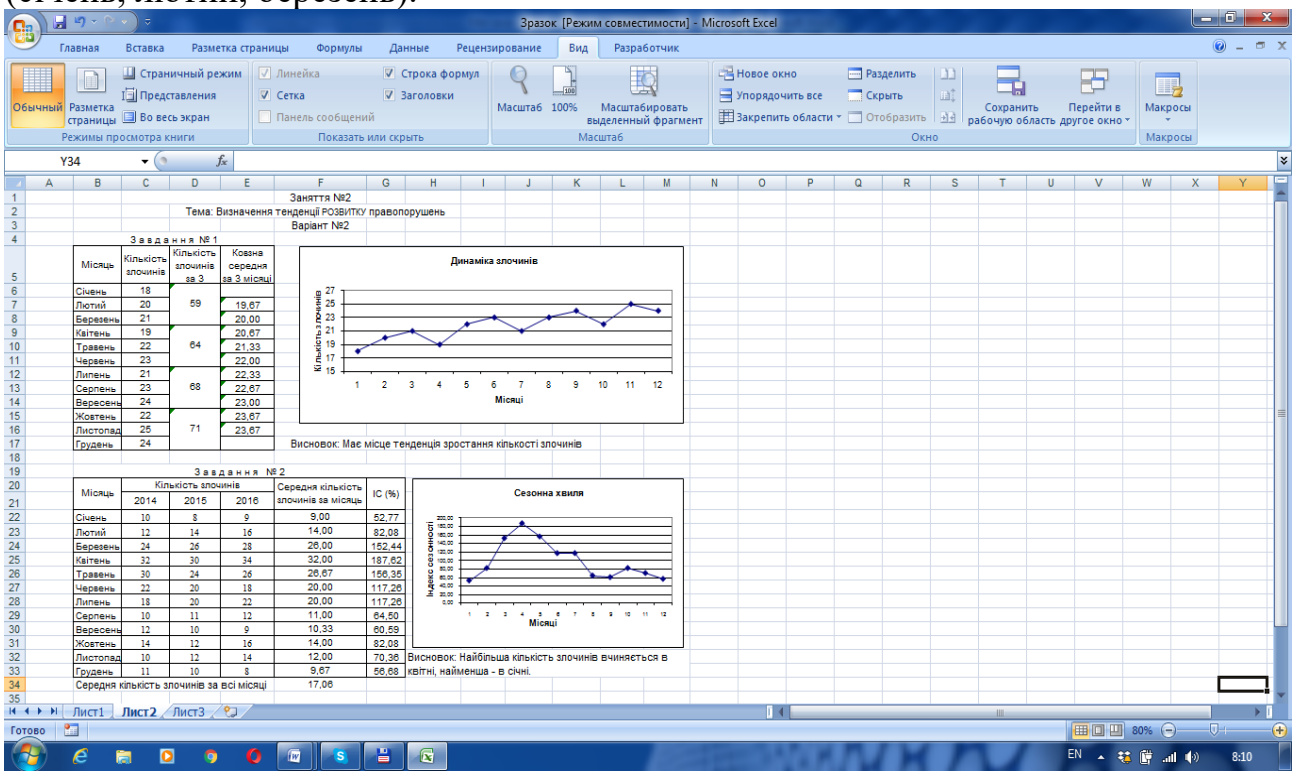
***Завдання №1*.** Визначить тенденцію зміни кількості зареєстрованих злочинів, використовуючи вирівнювання ряду динаміки методом укрупнення інтервалів часу і методом обчислення ковзної середньої. Для цього:

2.1. Вкажіть кількість злочинів за кожний місяць.

2.2. В комірках D7, D10, D13, D16 визначить кількість злочинів за 3 місяці (**укрупніть інтервали**), використовуючи функцію **СУММ**.

2.3. В комірках E7:E16 розрахуйте значення ***ковзної середньої*** за три місяці, використовуючи функцію **СРЗНАЧ**. Для цього:

2.3.1. В комірці E7 визначіть середню кількість злочинів за перші 3 місяці (січень, лютий, березень).



2.3.2. В комірці E8 визначіть середню кількість злочинів за 3 місяці, починаючи з 2-го місяця (лютий, березень, квітень).

2.3.3. В комірці E9 визначіть середню кількість злочинів за 3 місяці починаючи з 3-го місяця (березень, квітень, травень) і т.д..

Зробіть висновок щодо *тенденції* зміни кількості зареєстрованих злочинів.

3. Побудуйте графік “Динаміка злочинів”, який характеризує зміну кількості злочинів по місяцях.

Для цього:

3.1. Виділіть комірки **C6:C17** (кількість злочинів).

3.2. Виберіть у меню команду **Вставка**, потім в області **Диаграмма** натисніть піктограму **Графік** і виберіть діаграму **График с маркерами**.

3.3. Виберіть у меню вкладку **Макет**, у вікні **Подписи** натисніть **Название диаграммы,** вкажіть **Над диаграммой,** в заголовку **Название диаграммы** вкажіть **Динаміка злочинів.**

3.4. У вікні **Подписи** натисніть вікно **Название осей.**

3.5. Натисніть команду **Название основной горизонтальной оси**, після цього – **Название под осью**. В полі **Название оси** вкажіть **Місяці**.

3.6. У вікні **Подписи** натисніть вікно **Название осей**.

3.7. У вікні **Название осей** натисніть вікно **Название основной вертикальной оси**, у відкритому вікні вкажіть **Повёрнутое название**.

3.8. У відкритому вікні **Название оси** вкажіть **Кількість злочинів**.

3.9. У вікні **Подписи** натисніть вікно **Легенда**. У відкритому вікні вкажіть **Нет** (Не добавлять легенду).

3.10. Отриману діаграму встановіть у необхідне місце.

***Завдання №2*.** Визначить сезонні коливання злочинів, розрахувавши індекси сезонності. Для цього:

4.1. Вкажіть кількість злочинів за кожний місяць в 2014, 2015, 2016 роках.

4.2. В комірках F22:F33 розрахуйте середню кількість злочинів в кожному місяці за 3 роки ($\bar{y}_м$), використовуючи функцію **СРЗНАЧ**.

4.3. В комірці F34 розрахуйте середню кількість злочинів за всі місяці за 3 роки ($\bar{y}_{заг}$), використовуючи функцію **СРЗНАЧ**.

4.4. В комірках G22:G33 розрахуйте **індекси сезонності** за кожний місяць, використовуючи формулу $IC = \frac{\bar{y}_м}{\bar{y}_{заг}} \cdot 100\%$.

4.5. Вивчивши отримані індекси сезонності, зробіть висновок: у якому місяці вчиняється найбільша кількість злочинів, а в якому - найменша.

5. Таким же чином, як і попередній, побудуйте графік "**Сезонна хвиля**", який характеризує зміну індексів сезонності по місяцях.

Тема № 4. Статистичне дослідження великих даних.

Практичне заняття 4. Прогнозування стану злочинності.

Навчальна мета заняття: навчитися визначати прогнозовану кількість злочинів, використовуючи функції Microsoft Excel.

Час проведення: 2 год.

Навчальні питання:

1. Поняття ряду динаміки, види рядів динаміки.
2. Прогнозування значень статистичного показника.
3. Показники варіації ознаки.

Література:

Основна.

1. Інформаційні технології у правоохоронній діяльності. Частина 1: Високотехнологічні тренди у правоохоронній сфері зарубіжних країн: навч. посіб. / Харків. Нац. Ун-т внутр. Справ; [В.М. Струков, Д.В. Узлов, Ю.В. Гнусов та ін.] ; за заг. ред. канд. техн. наук, доц. В.М. Струкова. Харків : ТОВ «ДІСА ПЛЮС», 2020. 276 с.

2. Кубрак В. П. Правова статистика: навч. посіб. / В. П. Кубрак. – МВС України, Харк. нац. ун-т внутр.справ. – Харків: 2017. – 194 с.

3. Конспект лекцій.

Додаткова.

4. Зацеркляний М.М. Інформаційні технології у правозастосовній діяльності: Навч. посібник / М.М. Зацеркляний, В.М. Струков. : Х.: ТОВ „Східно-регіональний центр гуманітарно-освітніх ініціатив”; 2010. 332 с.

5. Зацеркляний М.М. Основи комп’ютерних технологій для економістів: Навч. посібник / М.М. Зацеркляний, О.Ф. Мельников, В.М. Струков. – К.: ВД „Професіонал”, 2006 р. – 672 с.

Інформаційні ресурси в Інтернеті.

6. Gartner: Топ-10 стратегічних трендів розвитку технологій у 2020 році: сайт. URL: https://ain.ua/2018/10/26/gartner-top-10-trendov-razvitiya-texnologij/ (дата звернення: 25.10.2019).

План проведення заняття.

Завдання: Відповідно до виданого Вам варіанта:

1) в *завданні №1* визначить прогнозовану кількість злочинів в 2017 і 2018 роках методом середнього абсолютного приросту;

2) в *завданні №2* визначить прогнозовану кількість злочинів в 2017 і 2018 роках методом середнього темпу зростання;

3) в *завданні №3* визначить прогнозовану кількість злочинів в 2017 і 2018 роках методом аналітичного вирівнювання ряду динаміки.

Порядок виконання.

1. Запустіть програму Excel.

2. Виконайте *завдання №1.* Визначить прогнозовану кількість злочинів в 2017 і 2018 роках методом середнього абсолютного приросту. Для цього:

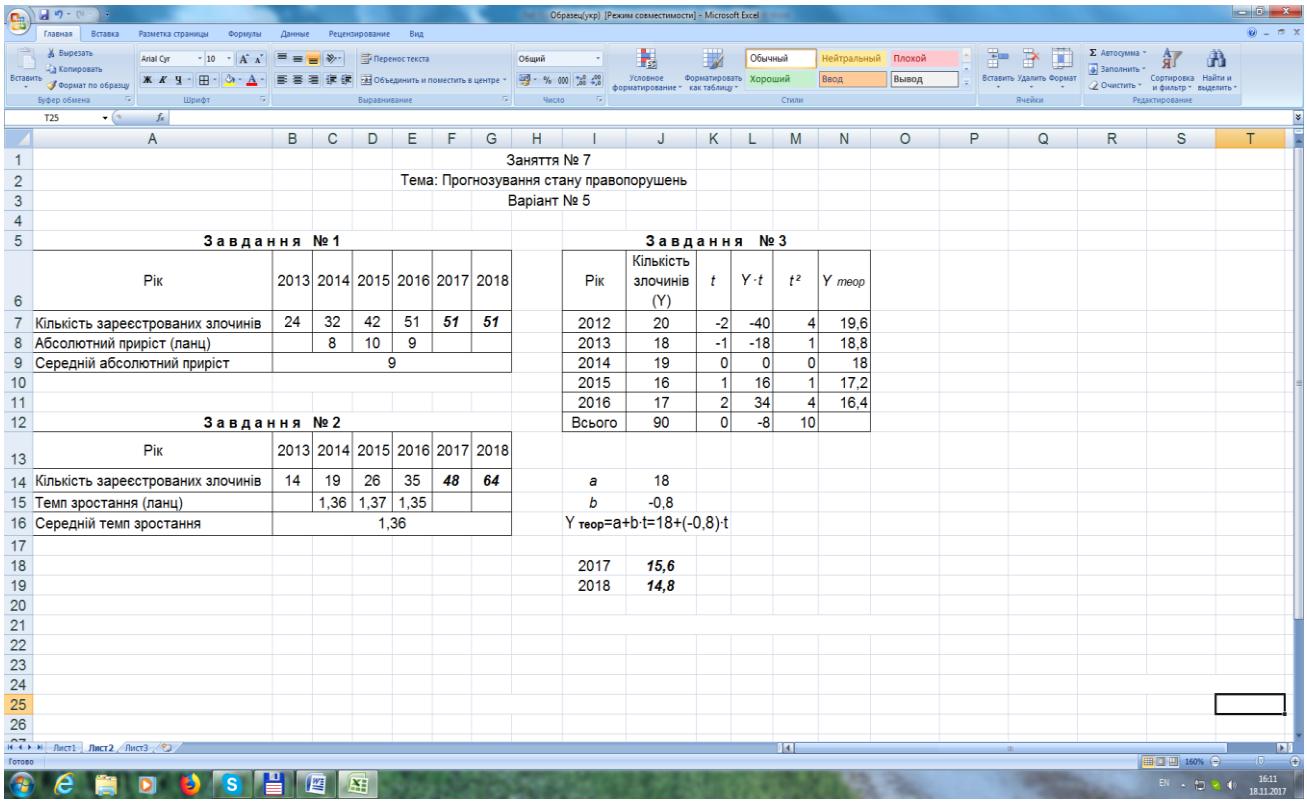
2.1. В комірки B6:G6 введіть роки.

2.4. В комірки B7:E7 введіть кількість зареєстрованих злочинів, зазначених у Вашому варіанті.

2.5. В комірках C8:E8 визначить абсолютний приріст ланцюговим способом в 2014, 2015 і 2016 роках.

2.6. В комірці B9 визначить середній абсолютний приріст $(\overline{A})$ за 2013-2016 роки.

2.7. В комірках F7 і G7 визначить прогнозовану кількість злочинів в 2017 і 2018 роках по формулі $Y_{n+t} = Y_n + \overline{A} \cdot t$, де: Y_n – останній рівень ряду динаміки, Y_{n+t} – прогнозований рівень, t – строк прогнозу (в 2017 році – 1, в 2018 році – 2).



Заняття № 7
Тема: Прогнозування стану правопорушень
Варіант № 5

Завдання № 1

Рік	2013	2014	2015	2016	2017	2018
Кількість зареєстрованих злочинів	24	32	42	51	***51***	***51***
Абсолютний приріст (ланц)		8	10	9		
Середній абсолютний приріст	9					

Завдання № 2

Рік	2013	2014	2015	2016	2017	2018
Кількість зареєстрованих злочинів	14	19	26	35	***48***	***64***
Темп зростання (ланц)		1,36	1,37	1,35		
Середній темп зростання	1,36					

Завдання № 3

Рік	Кількість злочинів (Y)	*t*	*Y·t*	t^2	*Y теор*
2012	20	-2	-40	4	19,6
2013	18	-1	-18	1	18,8
2014	19	0	0	0	18
2015	16	1	16	1	17,2
2016	17	2	34	4	16,4
Всього	90	0	-8	10	

a	18
b	-0,8

Y теор=a+b·t=18+(-0,8)·t

2017	***15,6***
2018	***14,8***

3. Виконайте *завдання №2.* Визначить прогнозовану кількість злочинів в 2017 і 2018 роках методом середнього темпу зростання. Для цього:

3.1.В комірки B13:G13 введіть роки.

3.2.В комірки B14:-E14 введіть кількість зареєстрованих злочинів, зазначених у Вашому варіанті.

3.3.В комірках C15:E15 визначить темп зростання ланцюговим способом в 2014, 2015 і 2016 роках.

3.4.В комірці B16 визначить середній темп зростання ($\overline{T}_{зр}$) за формулою: =(Тзр$_2$*Тзр$_3$*Тзр$_4$)^(1/3).

3.5.В комірках F14 і G14 визначить прогнозовану кількість злочинів в 2017 і 2018 роках по формулі $Y_{n+t} = Y_n \cdot \overline{T}_{зр}{}^t$, де: Y_n – останній рівень ряду динаміки, Y_{n+t} – прогнозований рівень, *t* – строк прогнозу (в 2017 році – 1, в 2018 році – 2).

4. Виконайте *завдання №3.* Визначить прогнозовану кількість злочинів в 2017 і 2018 роках методом аналітичного вирівнювання ряду динаміки. Для цього:

4.1. В комірки I6:N6 введіть найменування показників.

4.2. В комірки I7:I11 введіть роки.

4.3. В комірки J7:J11 введіть кількість зареєстрованих злочинів, зазначених у Вашому варіанті.

4.4. Зміна рівнів ряду динаміки апроксимується прямою лінією ($Y_{теор} = a + b \cdot t$). Визначить рівняння *конкретної* прямої лінії, що характеризує зміну рівнів *Вашого* ряду динаміки. Для цього:

4.4.1. Визначить ***a*** і ***b***. Для цього використовуємо умовний масштаб часу. Середній рівень (комірка **К9**) приймаємо за **0**, тоді вищестоящі рівні (комірки **К8**

і **К7**) будуть позначатися зі знаком *мінус* **(-1,-2**), а нижчестоящі (комірки **К10** і **К11**) – зі знаком *плюс* **(1, 2**).

4.4.2. Розрахуйте в комірках L7:L11 значення ($Y \cdot t$), а в комірках M7:M11 значення **t^2**.

4.4.3. Розрахуйте в комірках L12 і M12 суми значень ($Y \cdot t$) і **t^2** (у прикладі вони дорівнюють -8 і 10).

4.4.4. Розрахуйте в комірках J12 загальну кількість зареєстрованих злочинів.

4.4.5. Розрахуйте в комірках J10 і J11 значення ***a*** і ***b*** за формулами: $a = \frac{\sum Y}{n}$

$b = \frac{\sum (Y \cdot t)}{\sum t^2}$

Вони дорівнюють: ***a* = 18 *b* = -0,8.**

4.4.6. Знаючи ***a*** і ***b*** запишіть рівняння апроксимуючої прямої лінії для Вашого ряду динаміки (у прикладі воно виглядає так: $Y_{теор} = 18 + (-0{,}8) \cdot t$.

4.4.7. Підставляючи в отримане рівняння значення **t**, визначить в комірках N7:N11 значення **Yтеор**.

4.4.8. В комірках J18 і J19 визначить прогнозовану кількість злочинів в 2017 і 2018 роках. Для цього у формулі $Y_{теор} = 18 + (-0{,}8) \cdot t$ вкажіть значення ***t*=3** (для 2017 року) і ***t*=4** (для 2018 року).

Тема № 5. Алгоритми ієрархічної кластеризації.

Практичне заняття 5. Алгоритми ієрархічної кластеризації.

Навчальна мета заняття: відпрацювання практичних навичок використання ієрархічних алгоритмів кластерного аналізу в статистичній графічній системі STATGRAPHICS *Plus* for Windows».

Час проведення: 4 год.

Місце проведення - комп'ютерний клас

Навчальні питання:

1. Знайомство з основними алгоритмами кластерного аналізу.
2. Реалізація й графічне представлення ієрархічних агломеративних алгоритмів кластерного аналізу в системі STATGRAPHICS *Plus*;

Література:

Основна.

1. Aggarwal C.C. Data Mining. – Cham: Springer Ltd. Publ. Switzerland, 2015. – 734p.

2. Aggarwal C.C., Reddy C.K. Data Clustering. Algorithms and Applications.- New York:CRC Press, Tailor & Francik Group, 2014. – 648p.

3. Обзор методов кластеризации текстовой информации [Электронный ресурс] / К. М. Кириченко, М. Б. Герасимов - Электрон, текст, дан. - Режим доступа: www/ URL: http://www.dialog- 21.ru/Archive/2001/yolume2/2__26.htm - 10.12.2009 г. - Загл. с экрана.

4. Конспект лекцій.

Додаткова.

5. Інформаційні технології у правоохоронній діяльності. Частина 1: Високотехнологічні тренди у правоохоронній сфері зарубіжних країн: навч. посіб. / Харків. Нац. Ун-т внутр. Справ; [В.М. Струков, Д.В. Узлов, Ю.В. Гнусов та ін.] ; за заг. ред. канд. техн. наук, доц. В.М. Струкова. Харків : ТОВ «ДІСА ПЛЮС», 2020. 276 с.

6. Зацеркляний М.М. Інформаційні технології у правозастосовній діяльності: Навч. посібник / М.М. Зацеркляний, В.М. Струков. : Х.: ТОВ „Східно-регіональний центр гуманітарно-освітніх ініціатив"; 2010. 332 с.

Інформаційні ресурси в Інтернеті.

7. Gartner: Топ-10 стратегічних трендів розвитку технологій у 2020 році: сайт. URL: https://ain.ua/2018/10/26/gartner-top-10-trendov-razvitiya-texnologij/ (дата звернення: 25.10.2019).

План проведення заняття.

І. Порядок проведення вступу до заняття: ознайомлення здобувачів вищої освіти з навчальною метою заняття, навчальними питаннями та рекомендованою літературою.

ІІ. Порядок проведення основної частини заняття.

1.Завдання для виконання на практичному занятті:

1. Вивчити основні алгоритми кластерного аналізу.
2. Методами кластерного аналізу провести дослідження 16 інвестиційних фондів, що оперують цінними паперами[1] .
3. Реалізувати в системі STATGRAPHICS Plus ієрархічні агломеративні алгоритми кластерного аналізу.
4. Використовуючи інструмент системи **Statreporter**, скласти звіт по лабораторній роботі, що включає результати статистичного аналізу, побудовані таблиці й графіки.
5. Зберегти дані й результати роботи у вигляді файлу статистичного проекту **Statfolio**.

Порядок виконання завдання

Кластерний аналіз – це сукупність методів, що дозволяють провести угруповання багатомірних спостережень, кожне з яких описується набором вихідних змінних $X_1, X_2, ..., X_n$. Метою кластерного аналізу є створення груп схожих між собою об'єктів, які прийнято називати кластерами. Термін *кластер (cluster)* трактується як *згущення, скупчення, гроно, кисть, пучок, група.*

[1] Дюк В. Обработка данных на ПК в примерах. – СПб.: Питер, 1997 . – 240 с.

Методи кластерного аналізу можна розділити на дві більші групи ієрархічного групування: агломеративные (об'єднуючі) і дивизимні (поділяючі). Агломеративні методи послідовно поєднують окремі об'єкти в групи (кластери), а дивизимні методи розбивають групи на окремі об'єкти. У свою чергу кожний метод як об'єднуючого, так і розділяючого типу може бути реалізований за допомогою різних алгоритмів.

На першому кроці в агломеративних алгоритмах усі об'єкти вважаються окремими кластерами. Потім на кожному наступному кроці два найближчі кластери поєднуються в один. Кожне об'єднання зменшує число кластерів на одиницю. Виконання алгоритму завершується тоді, коли всі об'єкти поєднуються в один кластер. Найбільш підходящу розбивку вибирає найчастіше сам дослідник, якому надається дендрограма, що відображає результати групування об'єктів на всіх кроках алгоритму.

В STATGRAPHICS *Plus* for Windows реалізовано 7 видів ієрархічних агломеративних процедур і одна неієрархічна процедура кластерного аналізу: метод k-середніх. Ієрархічні процедури дозволяють простежити процес виділення угруповань і ілюструють вкладеність кластерів, що утворюються на різних етапах роботи.

Розглянемо задачу дослідження 16 інвестиційних фондів, що оперують цінними паперами, для оцінки їх стану. У якості змінних використовуються наступні характеристики (більшість із них описується в умовних одиницях): прибутковість за п'ятирічний період – змінна **Five_Yr,** ризик – змінна **Risk,** щорічний відсоток доходу (performance) – **Perf10, Perf11, Per12, Per13, Per14,** видаткова частина – змінна **Expence** і податкові рейтинги – змінна **Tax.**

Вихідні дані наведено в таблиці 2. У першому стовпці зазначене найменування фонду, а в останньому – рекомендації експертів по операціях із цінними паперами цих фондів.

Таблиця 2. Дані про інвестиційні фонди

Fund	Five_Yr	Risk	Perf10	Perf11	Perf12	Perf13	Perf14	Expence	Tax
Chip	16476	2	10	25	6	55	4	1,22	89
Contra	15476	2	–1	21	16	55	4	1,03	90
Destiny	14757	3	4	26	15	39	–3	0,70	69
Vista	15145	4	–1	20	13	71	–6	1,49	96
Berger	15596	5	–7	21	9	89	–6	1,70	95
Assett	13640	1	0	22	15	18	–6	1,33	85
Focus	14081	3	1	16	21	25	–6	0,85	75
Magellan	13827	3	–2	25	7	41	–5	0,96	73
Janus	13187	2	–1	11	7	43	–1	0,91	85
Mason	13029	4	1	12	11	35	–17	1,82	92
Gabelli	12301	3	–3	11	4	34	–2	1,41	80
Growth	11793	2	3	7	3	27	2	0,77	90
Janus	12441	4	–7	3	2	69	1	1,02	95
Capital	11728	4	–10	16	5	41	–16	0,97	68
Kemper	11386	4	–6	2	–2	67	4	1,09	86

Fund	Five_Yr	Risk	Perf10	Perf11	Perf12	Perf13	Perf14	Expence	Tax
20th Cent.	11258	4	–8	15	–4	32	0	1,00	60

1. Введемо наведені дані в електронну таблицю STATGRAPHICS і збережемо їх у файлі з іменем **Investfunddata** (рис. 1)**.**

InvestFundData.sf3

	Fund	Five_Yr	Risk	Perf10	Perf11	Perf12	Perf13	Perf14	Expence	Tax
1	Chip (1)	16476	2	10	25	6	55	4	1,22	89
2	Contra (2)	15476	2	-1	21	16	55	4	1,03	90
3	Destiny (3)	14757	3	4	26	15	39	-3	0,7	69
4	Vista (4)	15145	4	-1	20	13	71	-6	1,49	96
5	Berger (5)	15596	5	-7	21	9	89	-6	1,7	95
6	Assett (6)	13640	1	0	22	15	18	-6	1,33	85
7	Focus (7)	14081	3	1	16	21	25	-6	0,85	75
8	Magellan (8)	13827	3	-2	25	7	41	-5	0,96	73
9	Janus (9)	13187	2	-1	11	7	43	-1	0,91	85
10	Mason (10)	13029	4	1	12	11	35	-17	1,82	92
11	Gabelli (11)	12301	3	-3	11	4	34	-2	1,41	80
12	Growth (12)	11793	2	3	7	3	27	2	0,77	90
13	Janus (13)	12441	4	-7	3	2	69	1	1,02	95
14	Capital (14)	11728	4	-10	16	5	41	-16	0,97	68
15	Kemper(15)	11386	4	-6	2	-2	67	4	1,09	86
16	20th Cent. (16)	11258	4	-8	15	-4	32	0	1	60
17										
18										
19										

Рис. 1. Файл даних

2. Виберемо Special | Multivariate Methods | Cluster Analysis.

3. У вікні діалогу введення даних кластерного аналізу (рис. 2) запишемо змінні **Expence, Five_Yr, Perf10, Perf11, Perf12, Perf13, Perf14, Risk** і **Tax** у поле **Data**; характеристику **Fund** у поле **Point Labels**; поле даних **Select** залишимо порожнім.

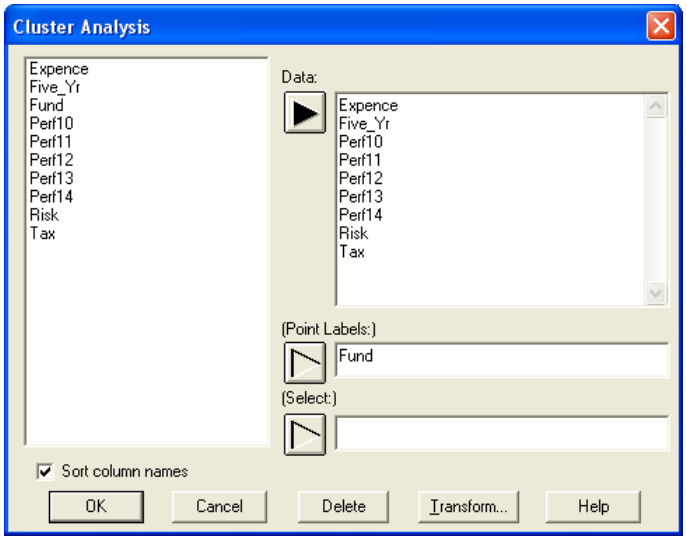


Рис. 2. Вікно діалогу введення даних кластерного аналізу

Після натискання клавіші **ОК** з'явиться вікно з первинним зведенням кластерного аналізу.

Команда **Analysis Options**... (викликається клацанням правої кнопки миші) панелі аналізу **Analysis Summary** дозволяє вибрати параметри алгоритму кластерного аналізу (рис. 3):

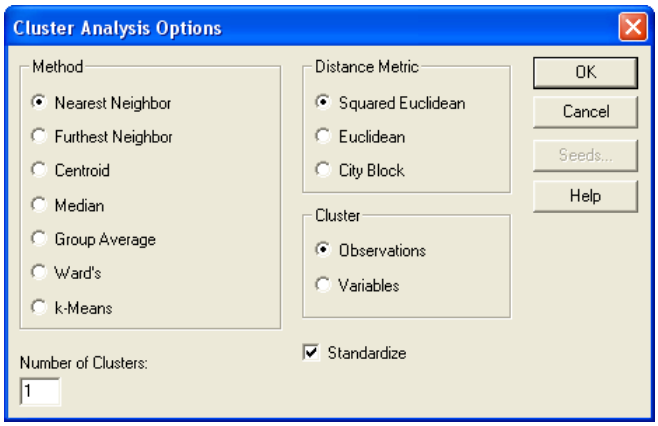


Рис. 3. Вікно вибору параметрів алгоритму кластерного аналізу

Розділ **Method** визначає метод об'єднання спостережень у кластери:
Nearest Neighbor – ступінь подібності двох кластерів визначається як відстань між двома найбільш схожими (найближчими) об'єктами цих кластерів (метод «близького сусіда»).

Furthest Neighbor – ступінь подібності двох кластерів визначається як відстань між двома найбільш несхожими (віддаленими) об'єктами цих кластерів (метод «далекого сусіда»).

Centroid – ступінь подібності двох кластерів визначається як відстань між центрами цих кластерів.

Median – ступінь подібності двох кластерів визначається як медіанна відстань від спостережень в одному кластері до спостережень в іншому кластері

Group Average – ступінь подібності двох кластерів визначається як середня відстань від спостережень в одному кластері до спостережень в іншому кластері.

Ward's – метод Уорда (Варда), при реалізації якого два кластери вважаються найбільше близькими, якщо при їхньому об'єднанні мінімізується приріст загальної дисперсії.

K-Means – неієрархічний метод К-Середніх, алгоритм якого припускає використання тільки вхідних значень змінних. Для початку процедури класифікації повинні бути задані *k* випадково обраних об'єктів, які будуть служити еталонами, тобто центрами кластерів.

Розділ **Distance Metric** визначає метод обчислення відстані між кластерами:

Squared Euclidean – квадрат евклідової відстані: $d_{ik} = \sum_{j=1}^{n} (x_{ij} - x_{kj})^2$.

Euclidean – евклідова відстань: $d_{ik} = \sqrt{\sum_{j=1}^{n} (x_{ij} - x_{kj})^2}$.

City Block – відстань city-block: $d_{ik} = \sum_{j=1}^{n} | x_{ij} - x_{kj} |$.

Розділ **Cluster** визначає схему кластеризації даних:

Obsevations – кластеризація проводиться за спостереженнями, тобто по рядках таблиці даних.

Variables – кластеризація проводиться по змінним, тобто по стовпцях таблиці даних.

Розділ **Number of Clusters** – кількість кластерів

Для кластеризації інвестиційних фондів за допомогою команди **Analysis Options**... (викликається клацанням правої кнопки миші) виберемо **Furthest Neighbor** – метод «далекого сусіда» об'єднання спостережень у кластери; метод обчислення відстані між кластерами – евклідова відстань **Euclidean**; інші параметри залишимо без зміни (рис. 4).

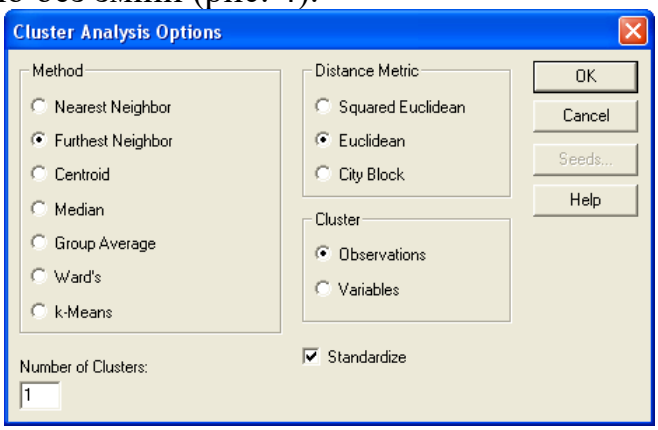


Рис. 4. Вікно параметрів кластеризації інвестиційних фондів

Після натискання клавіші **ОК** з'явиться вікно кластерного аналізу для обраних параметрів методу (рис. 5).

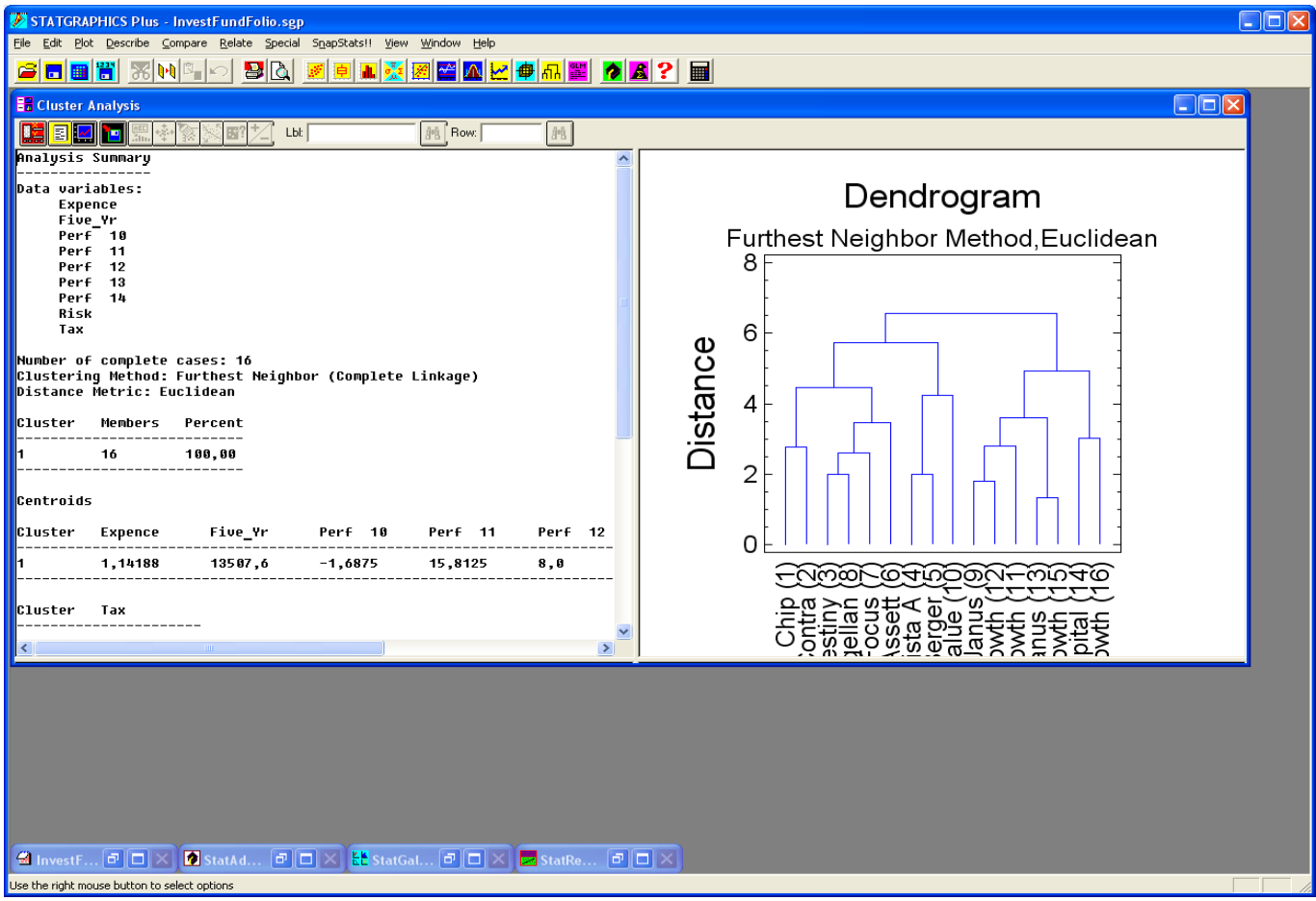


Рис. 5. Вікно кластеризации інвестиційних фондів

Дендрограма відображає ієрархічну структуру групування інвестиційних фондів. По вертикальній осі відкладена відстань для кожного кроку роботи агломеративного ієрархічного алгоритму кластеризації. На горизонтальній осі показані спостереження, скомбіновані відповідно до проведеного аналізу. На дендрограмі видні як мінімум три дерева (кластера) і імена об'єктів (інвестиційних фондів), що ввійшли у виділені кластери (рис. 5).

Для більш докладного розгляду кожного із кластерів за допомогою команди **Analysis Options**... (викликається клацанням правої кнопки миші) у вікні вибору параметрів алгоритму кластерного аналізу (рис. 5) установимо кількість кластерів **Number of Clusters** рівне 3.

Використовуючи кнопку **Tabular Options** – вибору табличних опцій, доповнимо результати процедури кластерного аналізу **Cluster Analysis** таблицями **Membership Table** (приналежність об'єктів кластеру) і **Agglomeration Schedule** (послідовність об'єднання об'єктів у кластери) (рис. 6).

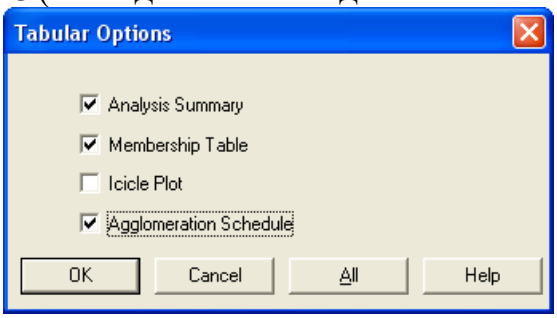


Рис. 6. Табличні опції процедури **Cluster Analysis**

Використовуючи кнопку **Graphics Options** – вибору графічних опцій, доповнимо результати процедури кластерного аналізу **Cluster Analysis** двовимірною діаграмою розсіювання **2D Scatterplot** (рис. 7).

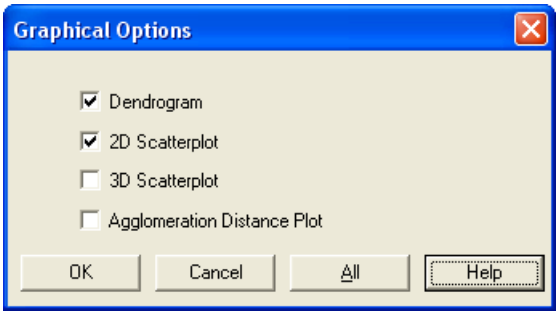


Рис. 7. Графічні опції процедури **Cluster Analysis**

Відповідно до введених змін будуть зроблені табличні й графічні перетворення (рис. 8).

У зведенні кластерного аналізу **Analysis Summary** вказуються: імена змінних, що брали участь в аналізі **Data variables**; кількість спостережень **Number of complete cases**; використаний метод кластерного аналізу **Clustering Method** і прийнята метрика **Distance Metric**. Далі у зведенні описуються: число кластерів **Cluster**, кількість об'єктів у кожному кластері **Members** і відсоток об'єктів, що належать кластеру **Percent**. У нижній частині зведення приводиться інформація про координати центрів кластерів **Centroids** по кожній зі змінних.

У таблиці **Membership Table** (приналежність об'єктів кластеру) описані обрані параметри кластерного аналізу й дається повний список усіх об'єктів, їх імена й номера кластерів, у які входять зазначені об'єкти.

У таблиці **Agglomeration Schedule** наведена послідовність об'єднання об'єктів у кластери.

Двовимірна діаграма розсіювання **2D Scatterplot** показує, як групуються досліджувані спостереження на площині двох змінні **Expence** і **Five_Yr** (рис. 9).

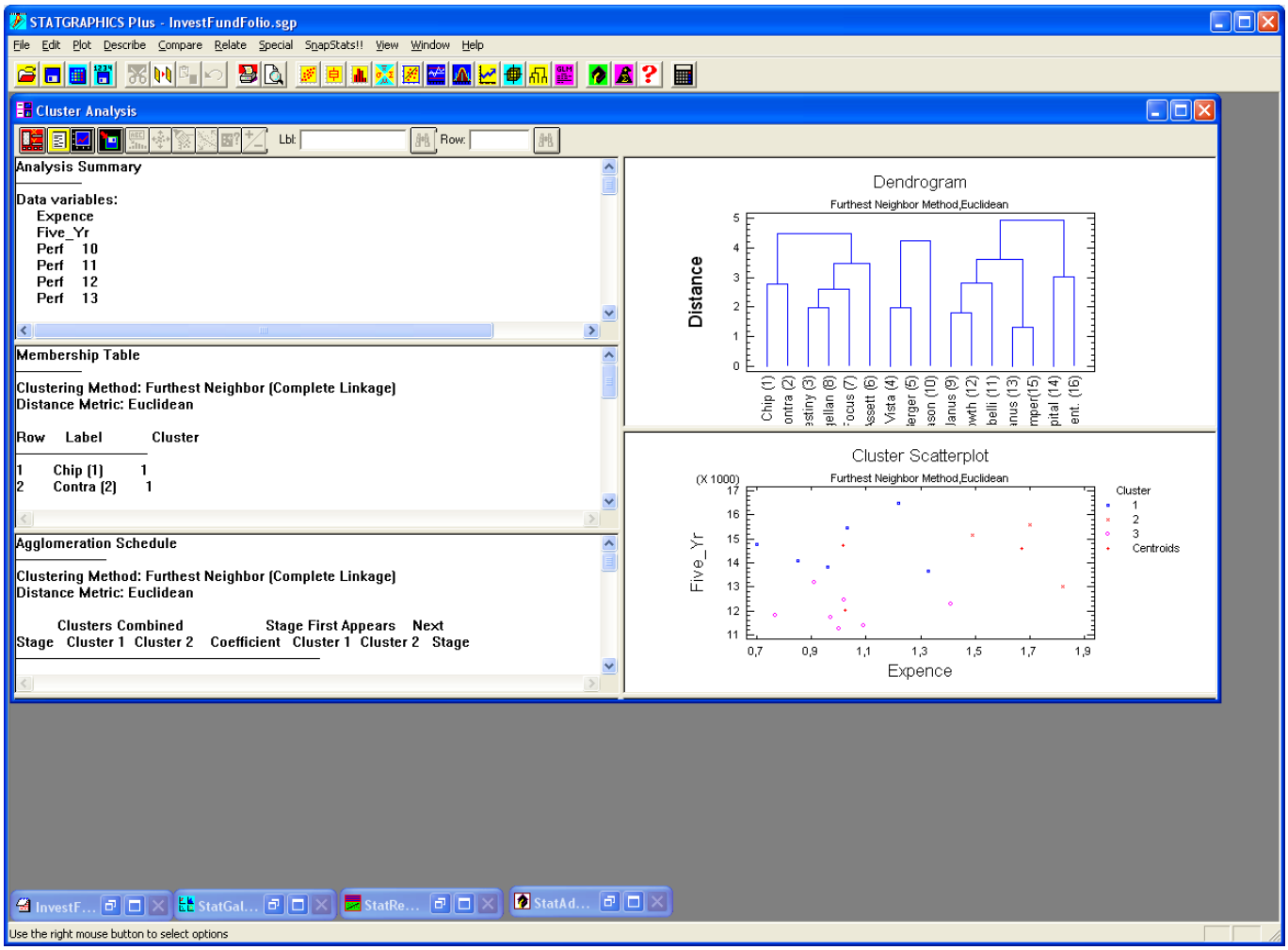


Рис. 8. Вікно процедури **Cluster Analysis** *угруповання фондів на три кластери*

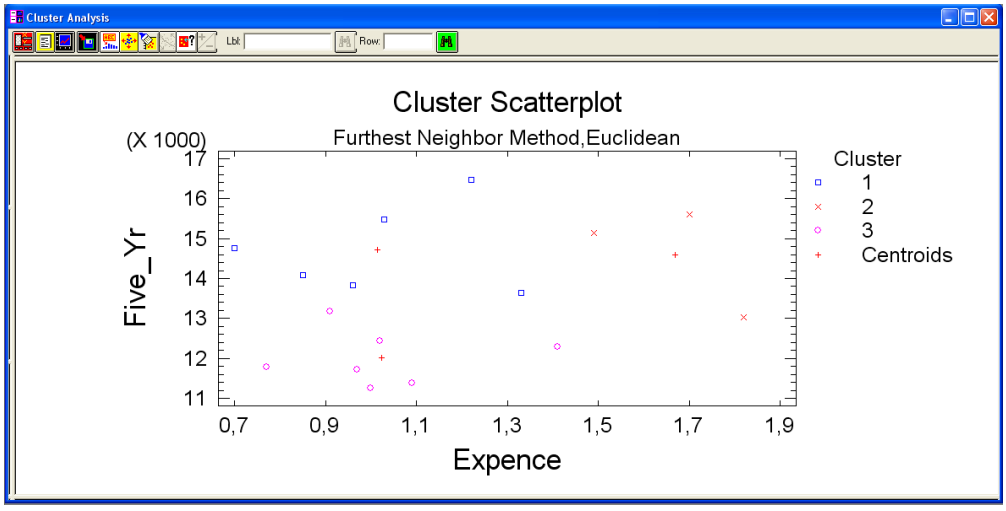


Рис. 9. Двовимірна діаграма розсіювання **2D Scatterplot**

Кожний кластер представлений на діаграмі власним символом і кольором. Із графіка випливає, що 1-й кластер має, низькі й середні відносні витрати **Expence** і досить високі доходи за п'ятирічний період. У кластері 2 спостерігаються найвищі витрати, але й середні й високі п'ятирічні доходи. У

кластері 3 низькі витрати супроводжуються й невисокими п'ятирічними доходами.

Для того щоб відобразити інші діаграми розсіювання, досить клацнути правою кнопкою миші й, у відповідному вікно діалогу, вибрати, що цікавлять пари змінних.

3. Складання звіту

Система STATGRAPHICS *Plus* містить спеціальний засіб складання звіту по проведеній статистичній роботі **Statreporter**, що дозволяє здійснити комбінування тексту й графіки.

Для збереження результатів статистичної обробки даних в **Statreporter** необхідно клацнути правою кнопкою миші в будь-якому місці обраної панелі аналізу й в, що з'явився контекстному меню вибрати пункт **Copy Analysis to Statreporter…**

4. Збереження статистичного проекту роботи

Для того щоб повторити всю схему й етапи проведеного статистичного аналізу для нового масиву даних система STATGRAPHICS *Plus* дозволяє зберігати результати роботи у вигляді спеціальних файлів статистичних проектів **Statfolio**. Для цього проводяться стандартні дії:

1. У головному меню виберіть команду **File | Save Statfolio As**;
2. У діалоговому вікні **Save Statfolio As** укажіть папку, у якій необхідно зберегти проект і ім'я файлу;
3. Натисніть кнопку **Save** (**Зберегти**).

Для проведення статистичного аналізу за схемою збереженого проекту **Statfolio** необхідно відкрити його командою **File | Open Statfolio**, а потім завантажити новий файл даних, виконавши команду **File | Open Datafile**. Усі задані статистичні процедури будуть автоматично перелічені з новими даними.

III. Порядок проведення заключної частини заняття

Для того щоб завершити роботу в системі STATGRAPHICS *Plus* можна виконати одне з наступних дій:

1. Вибрати команду **File | Exit Statgraphics**;
2. Натиснути комбінацію клавіш <**Alt+F4**>;
3. Натиснути кнопку закриття, яка перебуває в рядку заголовка вікна STATGRAPHICS *Plus*.

Тема № 4. Статистичне дослідження великих даних.

Практичне заняття 7-8. Дескриптивний аналіз великих даних.

Навчальна мета заняття: сформувати у студентів навички роботи в

статистичній графічній системі STATGRAPHICS *Plus* for Windows з варіаційними рядами, числовими характеристиками, інтервальним оцінюванням.

.

Час проведення: 4 год.

Навчальні питання:

1. Знайомство з основними властивостями системи і засобами роботи в STATGRAPHICS *Plus*; введення статистичних даних.
2. Побудова варіаційних рядів і їх графічних зображень у системі STATGRAPHICS *Plus*.
3. Обчислення числових характеристик статистичних розподілів.
4. Побудова інтервальних оцінок параметрів генеральної сукупності за вибірковим даними.

Література:

Основна.

1. Aggarwal C.C. Data Mining. – Cham: Springer Ltd. Publ. Switzerland, 2015. – 734p.

2. Aggarwal C.C., Reddy C.K. Data Clustering. Algorithms and Applications.- New York:CRC Press, Tailor & Francik Group, 2014. – 648p.

3. Обзор методов кластеризации текстовой информации [Электронный ресурс] / К. М. Кириченко, М. Б. Герасимов - Электрон, текст, дан. - Режим доступа: www/ URL: http://www.dialog- 21.ru/Archive/2001/yolume2/2__26.htm - 10.12.2009 г. - Загл. с экрана.

4. Конспект лекцій.

Додаткова.

5. Інформаційні технології у правоохоронній діяльності. Частина 1: Високотехнологічні тренди у правоохоронній сфері зарубіжних країн: навч. посіб. / Харків. Нац. Ун-т внутр. Справ; [В.М. Струков, Д.В. Узлов, Ю.В. Гнусов та ін.] ; за заг. ред. канд. техн. наук, доц. В.М. Струкова. Харків : ТОВ «ДІСА ПЛЮС», 2020. 276 с.

6. Зацеркляний М.М. Інформаційні технології у правозастосовній діяльності: Навч. посібник / М.М. Зацеркляний, В.М. Струков. : Х.: ТОВ „Східно-регіональний центр гуманітарно-освітніх ініціатив”; 2010. 332 с.

Інформаційні ресурси в Інтернеті.

7. Gartner: Топ-10 стратегічних трендів розвитку технологій у 2020 році: сайт. URL: https://ain.ua/2018/10/26/gartner-top-10-trendov-razvitiya-texnologij/ (дата звернення: 25.10.2019).

План проведення заняття:

I. Порядок проведення вступу до заняття: ознайомлення здобувачів вищої освіти з навчальною метою заняття, навчальними питаннями та рекомендованою літературою.

II. Порядок проведення основної частини заняття.

1. Завдання для виконання.

1. Вивчити основні характеристики системи STATGRAPHICS *Plus*.
2. Ввести значення змінної X для 50 одиниць досліджуваної сукупності: $x_1, x_2, \ldots, x_n$ (табл. 1).

Таблиця 1.

x_1	x_2	x_3	x_4	x_5
98	**100**	**98**	**100**	**101**
x_6	x_7	x_8	x_9	x_{10}
100	**102**	**101**	**101**	**102**
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
99	**101**	**101**	**103**	**99**
x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
100	**100**	**98**	**99**	**103**
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
98	**100**	**99**	**98**	**102**
x_{26}	x_{27}	x_{28}	x_{29}	x_{30}
101	**100**	**100**	**100**	**101**
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}
99	**101**	**103**	**97**	**102**
x_{36}	x_{37}	x_{38}	x_{39}	x_{40}
103	**99**	**100**	**100**	**99**
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}
101	**101**	**100**	**102**	**97**
x_{46}	x_{47}	x_{48}	x_{49}	x_{50}
100	**103**	**102**	**100**	**97**

3. Використовуючи значення змінної X, побудувати дискретний варіаційний ряд, стовпчикову й секторну діаграми.

4. Обчислити описові статистики (числові характеристики) змінної X: середнє, медіану, моду, дисперсію, середнє квадратичне відхилення, мінімум, максимум, розмах варіювання, нижній квартиль, верхній квартиль, коефіцієнт асиметрії, стандартизований коефіцієнт асиметрії, коефіцієнт ексцесу, стандартизований коефіцієнт ексцесу, коефіцієнт варіації.

5. За даними вибірки обчислити 1-й, 10-й, 25-й, 35-й, 50-й, 65-й, 75-й, 90-й, 99-й процентили.

6. За даними вибірки побудувати 95% і 99% довірчі інтервали для математичного очікування й середнього квадратичного відхилення.

7. За даними вибірки побудувати інтервальний варіаційний ряд і гістограму частот.

8. Використовуючи інструмент системи **Statreporter**, скласти звіт по лабораторній роботі, що включає результати статистичного аналізу, побудовані таблиці й графіки.

9. Зберегти дані й результати роботи у вигляді файлу статистичного проекту **Statfolio**.

2. Обробка даних у статистичній системі STATGRAPHICS Plus

STATGRAPHICS *Plus* – програмний пакет для статистичного аналізу даних, що включає більше 250 статистичних і системних процедур, що застосовуються в бізнесі, економіці, маркетингу, медицині, біології, соціології, психології й в інших областях.

У базовій системі функціонують наступні процедури:

- Меню **Describe** (опис даних) містить статистичні методи аналізу по одній і множині змінних, процедури добору розподілів, засоби табуляції даних;
- Меню **Compare** (порівняння даних) включає методи порівняння двох і більше вибірок даних, процедури одне– і багатофакторного дисперсійного аналізу;
- Меню **Relate** (відносини даних) містить процедури простого, поліноміального й множинного регресійного аналізу.

Для розширення можливостей системи пропонуються додаткові модулі, ініціалізація яких здійснюється через меню **Special**. До них відносяться:

- Модуль «**Контроль качества**» призначений для оцінки ефективності всіх ланок виробничого процесу й формування відповідних карт.
- Модуль «**Планирование эксперимента**» допомагає сформулювати критерій оптимальності плану експерименту, підібрати найкращий план, організувати збір і обробку необхідної інформації. У модулі пропонуються ефективні способи спрощення й інтеграції знань про досліджуваний процес.
- Модуль «**Анализ временных рядов**» містить описові методи, процедури згладжування рядів, сезонної декомпозиції й прогнозування.
- Модуль «**Многомерные методы**» призначений для вивчення взаємин множини факторів (змінних). Для цього в модулі функціонує п'ять процедур, що забезпечують проведення **Кластерного аналізу**, аналізу по методу **Головних компонентів**, **Факторного**, **Дискримінантного** й **Канонічного кореляційного** аналізу.
- **Розширений регресійний аналіз**, крім базисних процедур регресійного аналізу, включає різні калібровані моделі, процедури порівняння ліній регресії, відбору найкращих регресійних моделей, нелінійну множинну регресію, ридж–регресію й логістичну регресію.

Statfolio – статистичний проект

В STATGRAPHICS *Plus* реалізований засіб створення статистичних проектів для автоматизації й збереження результатів роботи. Усю процедуру

аналізу даних (обрані методи, параметри статистичних методів, види графічних відображень результатів аналізу, табличні форми, коментарі й т.п.), можна зберегти у вигляді файлу **Statfolio**. Якщо виникає необхідність в обробці іншої множини даних за складеною схемою аналізу, потрібно завантажити новий файл даних. Результати аналізу, таблиці й графіки будуть перелічені автоматично.

Statadvisor – статистична консультація

У програмі STATGRAPHICS *Plus* реалізований інструмент **Statadvisor** інтерпретації результатів аналізу даних; обговорення можливих недоліків і проблем методів аналізу; визначення статистичної значимості отриманих результатів.

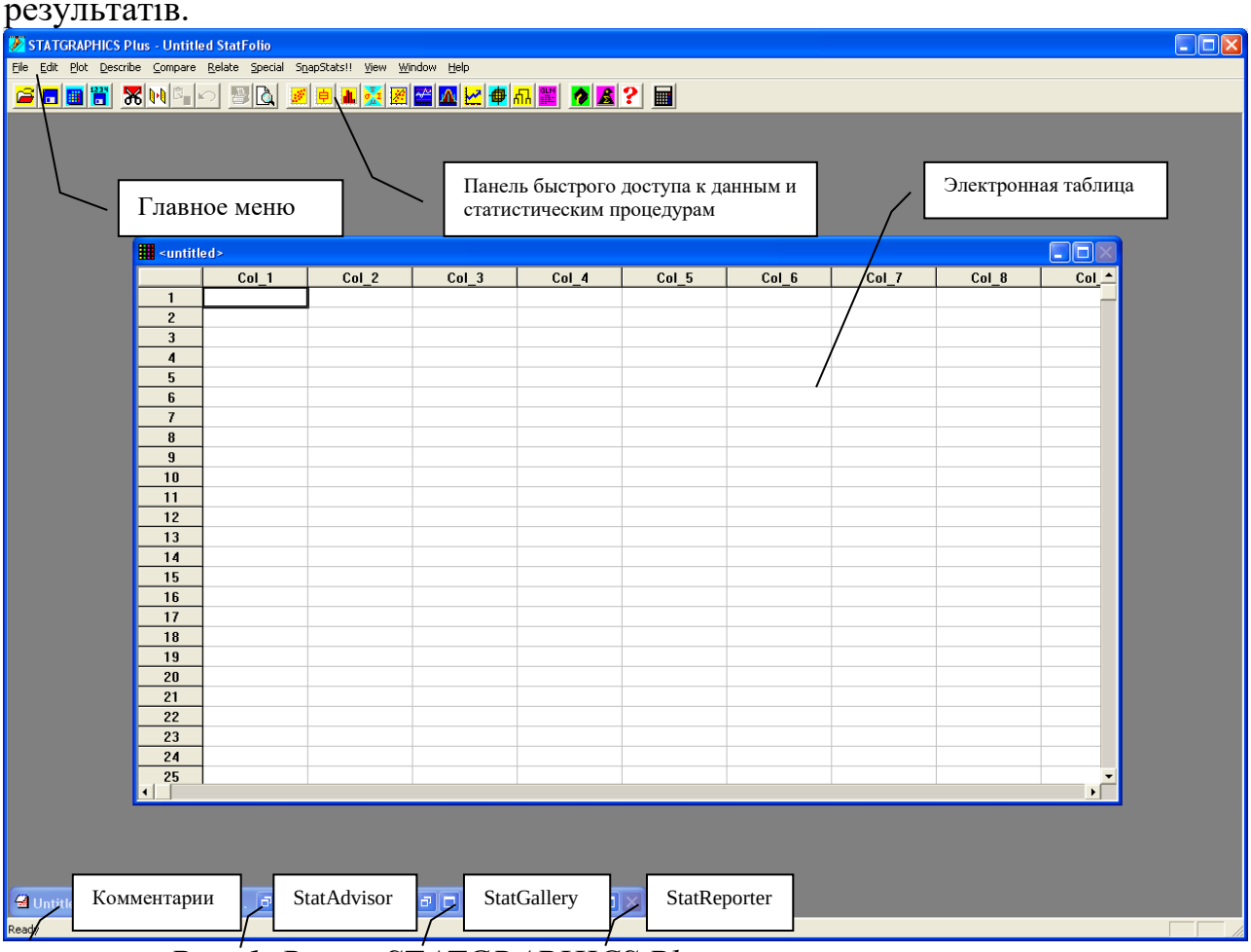


Рис. 1. Вікно STATGRAPHICS Plus і його основні елементи

Statgallery – інструмент для комбінування тексту й графіки

Інструмент **Statgallery** дозволяє зберігати результати аналізу даних у вигляді метафайла для архівації, перегляду і друку.

Statreporter – інструмент для створення звітів

Statreporter – «проміжний» інструмент між блокнотом і «повним» текстовим процесором, дозволяє поєднувати звіти, створені з табличних опцій, графіків, заміток і інтерпретацій з Statadvisor.

Вид робочого вікна STATGRAPHICS *Plus* представлений на рис. 1.

Набір кнопок у панелі швидкого доступу призначений для відкриття й збереження статистичних проектів **Statfolio**, файлів даних, для роботи з буфером обміну, для виведення результатів статистичного аналізу на друк, а також для виклику деяких статистичних і графічних процедур.

У нижній частині вікна розташований набір піктограм, зв'язаних з наступними операціями: **Untitled Comments, Statadvisor, Statgallery, Statreporter.**

3. Введення даних

1. Після запуску STATGRAPHICS *Plus* відкривається вікно **Untitled** електронної таблиці (рис. 1). Ця таблиця організована таким чином, що її рядки відповідають спостереженням (експериментальним, вибірковим даним), а стовпці – ознакам (змінним). Робота з нею аналогічна роботі з іншими відомими електронними таблицями для Windows, такими як Microsoft Excel, Statistica, SPSS, Lotus і ін.

2. Для завдання імені й типу змінних (ознак) виконайте наступні дії:

* клацніть лівою клавішею миші на імені змінюваної змінної (наприклад, на **Col_1**);

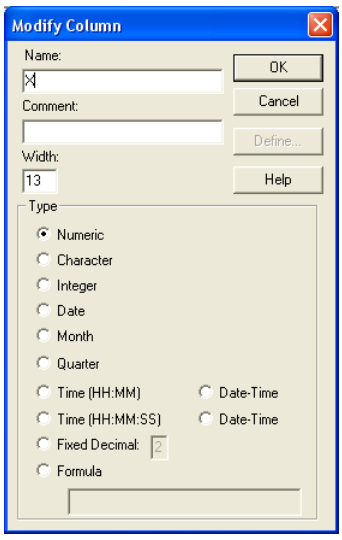


Рис. 2. Вікно модифікації змінної

* клацніть правою кнопкою миші;
* у контекстному меню, що з'явився, виберіть **команду Modify** Column;
* у вікні діалогу **Modify Column** (рис. 2) заповніть поля: **Name** – ім'я змінної, (наприклад, **X**); **Comment** – коментар (при необхідності); у поле **Type** відзначте необхідний тип змінної.

*натисніть **OK**;

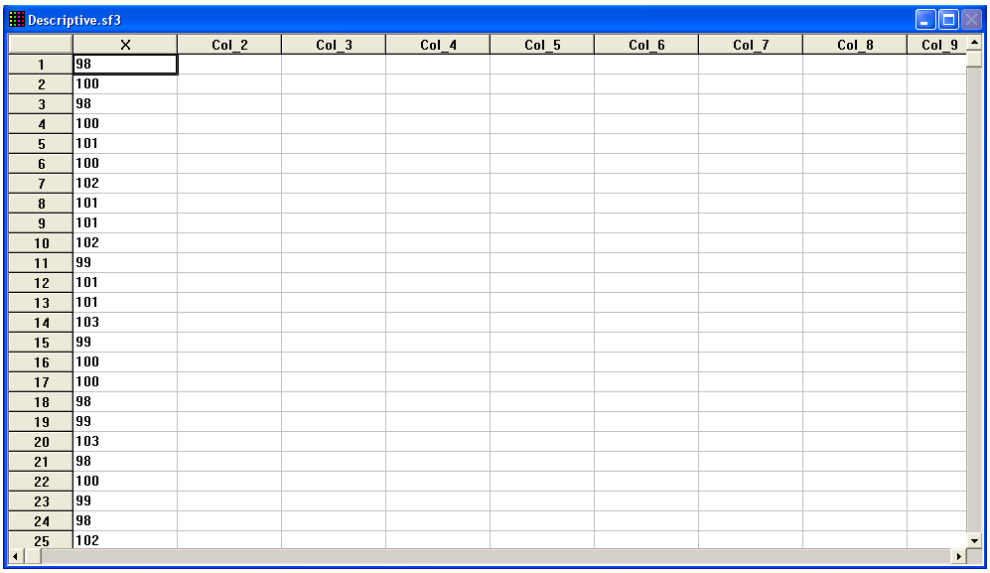
*скасуйте виділення стовпчика, установивши покажчик миші поза границею стовпчика й клацнувши лівою кнопкою.

3. Введіть значення змінної.

4. Після заповнення таблиці необхідно зберегти файл даних, для чого виберіть команду **File | Save | Save Data File** (або натисніть на піктограму  на панелі швидкого доступу)**;** введіть ім'я файлу (наприклад, **Descriptive**) і

натисніть **OK**. Після цієї операції в заголовку таблиці з'явиться зазначене ім'я, яке буде використовуватися у подальшому аналізі (рис. 3).

5. Збережіть проект аналізу даних, для чого виберіть команду **File | Save | Save Statfolio** (або натисніть на піктограму  на панелі швидкого доступу)**;** введіть ім'я проекту (наприклад, **Descriptivefolio**) і натисніть **OK**. Після цієї операції в заголовку робочого вікна STATGRAPHICS *Plus* з'явиться зазначене ім'я проекту.



	X	Col_2	Col_3	Col_4	Col_5	Col_6	Col_7	Col_8	Col_9
1	98								
2	100								
3	98								
4	100								
5	101								
6	100								
7	102								
8	101								
9	101								
10	102								
11	99								
12	101								
13	101								
14	103								
15	99								
16	100								
17	100								
18	98								
19	99								
20	103								
21	98								
22	100								
23	99								
24	98								
25	102								

Рис. 3. Файл даних

4. Побудова й графічне зображення дискретного варіаційного ряду

1. У головному меню виберіть команду **Describe | Categorical Data | Tabulation**.

2. У вікні, що з'явилося, вибору змінних **процедури** Tabulation (рис. 4) відзначте *необхідну* змінну.

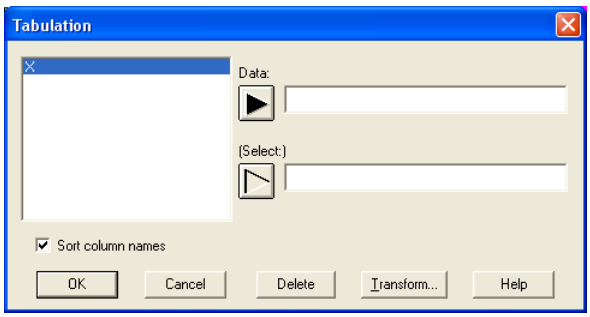


Рис. 4. Вікно вибору змінних процедури **Tabulation**

3. Внесіть ім'я змінної в поле **Data**, клацнувши лівою кнопкою миші на кнопці  .

4. Натисніть клавішу **OK**.

5. На екрані з'явиться вікно **Tabulation, що** містить такі види аналізу

обраної змінної, як:

1) **Analysis Summary** – узагальнений аналіз (**Data variable**: – аналізована змінна; **Number of observations**: – число спостережень; **Number of unique values**: – число різних значень спостережень (число варіант));

2) **Frequency Table** – таблиця частот. тобто варіаційні ряди частот (**Class** – порядковий номер варіанти; **Value** – варіанта; **Frequency** – частота; **Relative Frequency** – відносна частота; **Cumulative Frequency** – накопичена частота; **Cum. Rel. Frequency** – накопичена відносна частота);

3) **Barchart** – стовпчикова (смугова) діаграма.

4) **Piechart** – секторна діаграма.

У верхньому рядку вікна **Tabulation** розташовані стандартні кнопки, які з'являються у вікні будь-якої статистичної процедури (рис. 5):

Input Dialog – зміна вхідних даних;

Tabular Options – вибір табличних опцій;

Graphics Options – вибір графічних опцій;

Save results – збереження результатів аналізу у файлі даних.

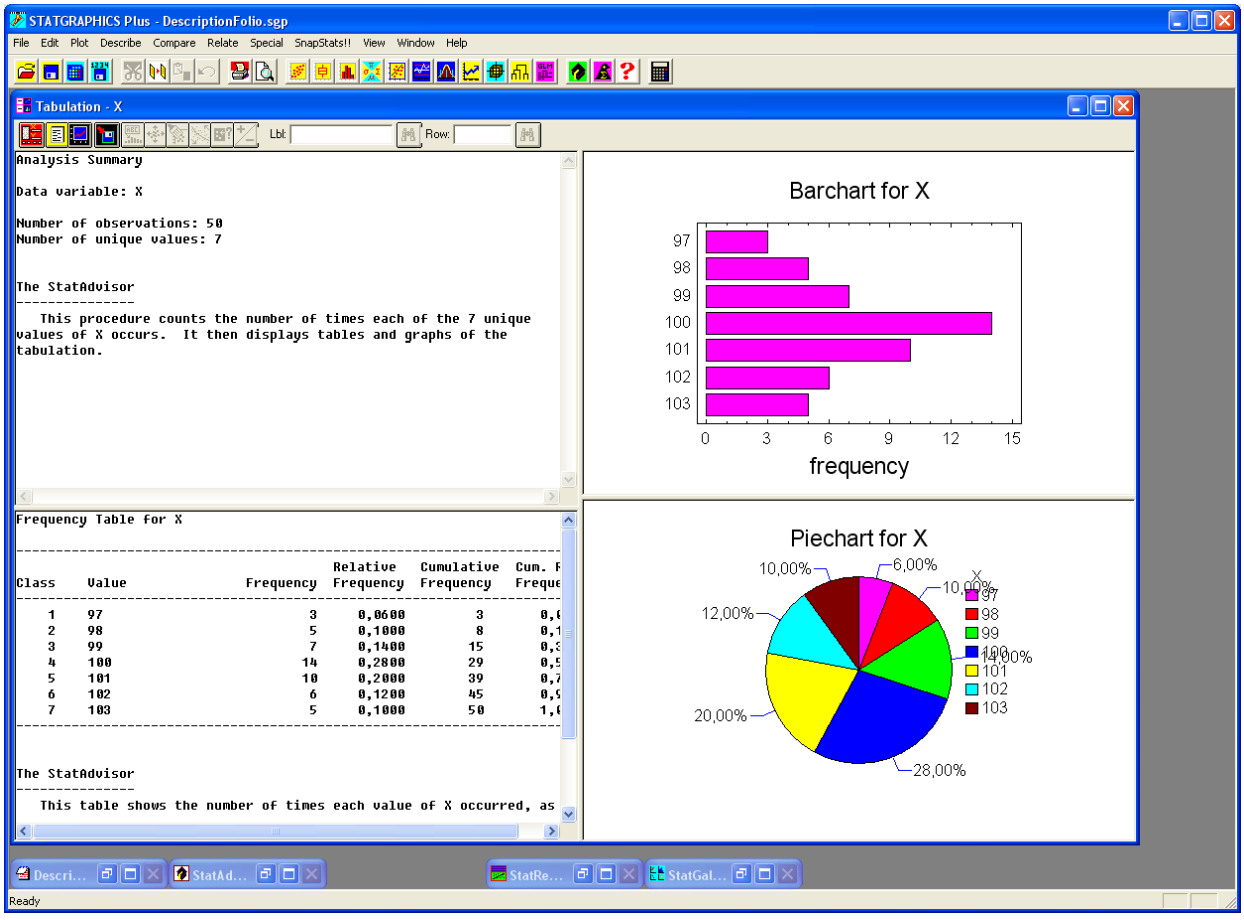


Рис. 5. Вікно процедури **Tabulation**

Подвійним клацанням лівої кнопки миші по обраній панелі аналізу можна

розкрити його для перегляду.

Клацання правої кнопки миші на панелі аналізу викликає контекстне меню опцій, пункти якого залежать від типу панелі. Наприклад, контекстне меню стовпчикової діаграми **Barchart** (рис. 6) дозволяє одержати доступ до опцій панелі (**Pane Options**...), параметрам графіки (**Graphics Options**...), скопіювати графік у буфер обміну (**Copy**…), роздрукувати (**Print**…) або виконати перегляд графіка перед друком (**Print Preview**…), скопіювати панель у галерею (**Copy Pane to Gallery**…) або звіт (**Copy Analysis to Statreporter**…), зберегти графік (**Save Graph**…).

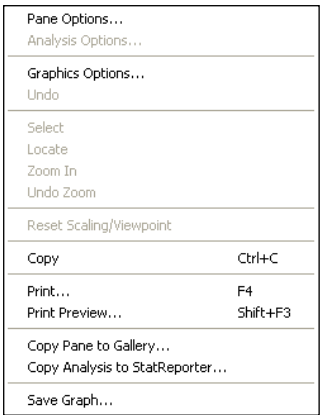


Рис. 6. Контекстне меню стовпчикової діаграми **Barchart**

Використовуючи пункт **Pane Options**... контекстного меню стовпчикової діаграми **Barchart** змінимо горизонтальний напрямок графіка на вертикальне (рис. 7):

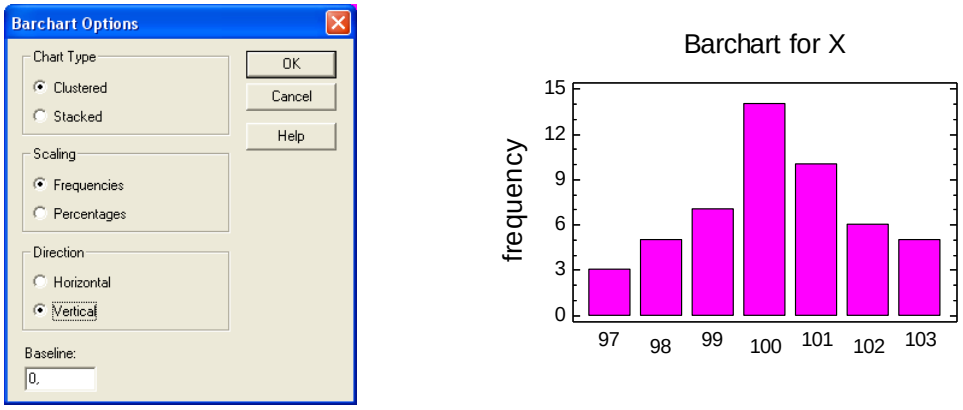


Рис. 7. Зміна напрямку столбиковой діаграми **Barchart**

5. Одномірний аналіз даних

Числові характеристики (сумарні статистики)

1. У головному меню виберіть команду **Describe | Numeric Data | One-Variable Analysis**.

2. У діалоговому вікні **One-Variable Analysis** відзначте *досліджувану* змінну.

3. Запишіть ім'я змінної в поле **Data**.

4. Натисніть клавішу **OK**.

5. На екрані з'явиться вікно **One-Variable Analysis** (рис. 8)**, що** містить такі види аналізу обраної змінної, як:

1) **Analysis Summary** – узагальнений аналіз (**Data variable**: – аналізована змінна; **Number of observations**: – число спостережень; **Number of unique values**: – число різних значень спостережень (число варіант)).

2) **Summary Statistics** – сумарні статистики.

За замовчуванням обчислюються наступні характеристики:

Average - середня арифметична;

Median - медіана;

Mode - мода;

Variance - дисперсія;

Std. Deviation - середнє квадратичне відхилення (стандартне відхилення);

Min. - мінімум;

Max. - максимум;

Range - розмах варіювання;

Lower Quartile - нижній квартиль;

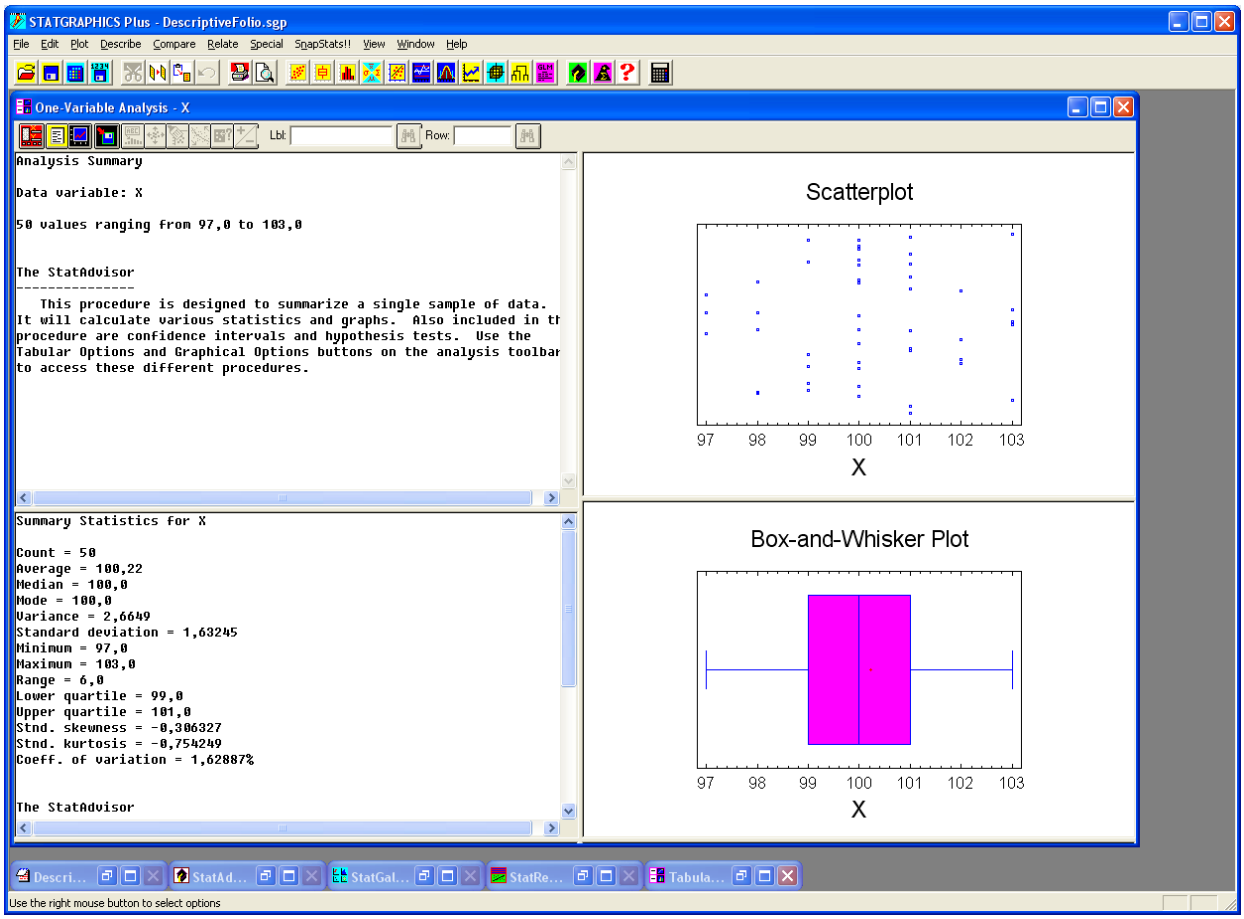


Рис. 8. Вікно процедури **One-Variable Analysis**

Upper Quartile - верхній квартиль;
Std. Skewness - стандартизований (нормований) коефіцієнт асиметрії;
Std. Kurtosis - стандартизований (нормований) коефіцієнт ексцесу;
Coeff. of Var. - коефіцієнт варіації;
3) **Scatterplot** – діаграма розсіювання.
4) **Box-and-Whisker Plot** – діаграма "Ящик з вусами" (діаграма розмаху).

Команда **Pane Options**... (викликається клацанням правої кнопки миші) панелі аналізу **Summary Statistics** дозволяє обчислити додаткові характеристики (рис. 9):

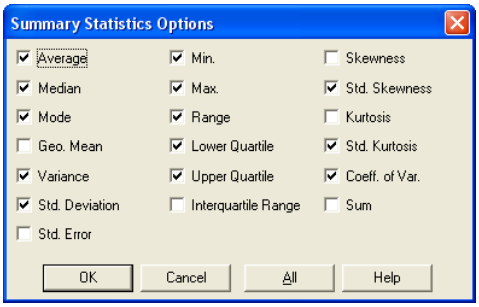


Рис. 9. Вікно вибору статистик панелі аналізу **Summary Statistics**

Geo. Mean - середнє геометричне;
Std. Error - стандартна помилка;
Interquartile Range - міжквартильний розмах;
Skewness - коефіцієнт асиметрії;

Kurtosis - коефіцієнт ексцесу;
Sum - сума елементів сукупності;

Використовуючи кнопку **Tabular Options** – вибору табличних опцій, доповнимо одномірний аналіз даних **One-Variable Analysis** процедурами обчислення процентилів, побудови інтервального варіаційного ряду й довірчих інтервалів (рис. 10).

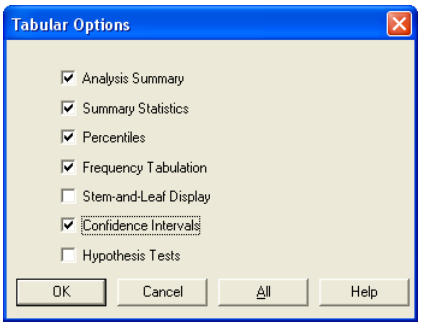


Рис. 10. Вікно вибору видів одномірного аналізу **One-Variable Analysis**

6. Побудова й графічна зображення інтервального варіаційного ряду

1. Для побудови інтервального варіаційного ряду за вибірковим даними натисніть кнопку вибору табличних опцій **Tabular Option** одномірного аналізу **One-Variable Analysis**.
2. У вікні, що з'явилося, відзначте **пункт Frequency** Tabulation (мал. 10).
3. Натисніть клавішу **OK**.
4. У панелі одномірного аналізу **One-Variable Analysis** з'явиться інтервальний варіаційний ряд:

Frequency Tabulation for X

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		96,0		0	0,0000	0	0,0000
1	96,0	97,1429	96,5714	3	0,0600	3	0,0600
2	97,1429	98,2857	97,7143	5	0,1000	8	0,1600
3	98,2857	99,4286	98,8571	7	0,1400	15	0,3000
4	99,4286	100,571	100,0	14	0,2800	29	0,5800
5	100,571	101,714	101,143	10	0,2000	39	0,7800
6	101,714	102,857	102,286	6	0,1200	45	0,9000
7	102,857	104,0	103,429	5	0,1000	50	1,0000
above	104,0			0	0,0000	50	1,0000

Mean = 100,22 Standard deviation = 1,63245

Таблиця варіаційного ряду містить наступні стовпці:
Class – Порядковий номер

Lower Limit – Нижня границя
Upper Limit – Верхня границя
Midpoint – Середина інтервалу
Frequency – Частота
Relative Frequency – Відносна частота
Cumulative Frequency – Накопичена частота
Cumulative Relative Frequency – Накопичена відносна частота
Mean – Середня арифметична
Standard deviation – Середнє квадратичне відхилення

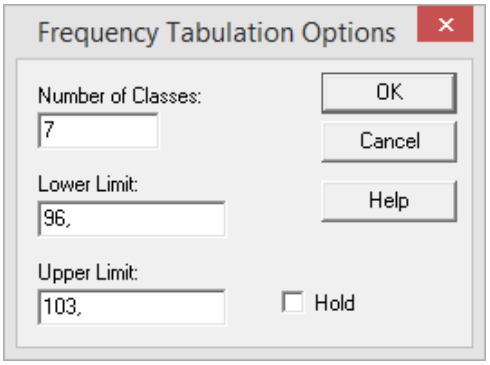


Рис. 12. Вікно зміни опцій процедури **Frequency Tabulation**

Для зміни параметрів варіаційного ряду клацніть правою кнопкою в будь-якій графі варіаційного ряду. У діалоговому меню, що з'явився, виберіть **пункт Pane** Options. З'явиться вікно **Frequency Tabulation Options** (рис. 12).

Заповнивши відповідні поля, можна змінити:

*****число інтервалів розбивки (**Number of Classes**);

*****нижню границю першого інтервалу розбивки (**Lower Limit**);

*****верхню границю останнього інтервалу розбивки (**Upper Limit**).

Frequency Tabulation for X

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		96,0		0	0,0000	0	0,0000
1	96,0	97,0	96,5	3	0,0600	3	0,0600
2	97,0	98,0	97,5	5	0,1000	8	0,1600
3	98,0	99,0	98,5	7	0,1400	15	0,3000
4	99,0	100,0	99,5	14	0,2800	29	0,5800
5	100,0	101,0	100,5	10	0,2000	39	0,7800
6	101,0	102,0	101,5	6	0,1200	45	0,9000
7	102,0	103,0	102,5	5	0,1000	50	1,0000
above	103,0			0	0,0000	50	1,0000

Mean = 100,22 Standard deviation = 1,63245

Для того щоб зафіксувати внесені зміни встановите прапорець **Hold.**

Використовуючи кнопку **Graphics Options** – вибір графічних опцій

(рис. 13), додамо до графічних панелей одномірного аналізу даних **One-Variable Analysis** панель **Frequency Histogram** з гістограмою частот (рис. 14).

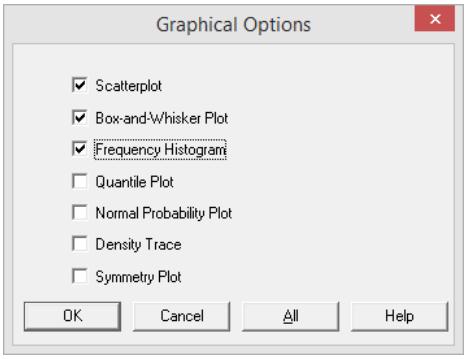


Рис. 13. Вікно вибору графічних опцій **Graphics Options**

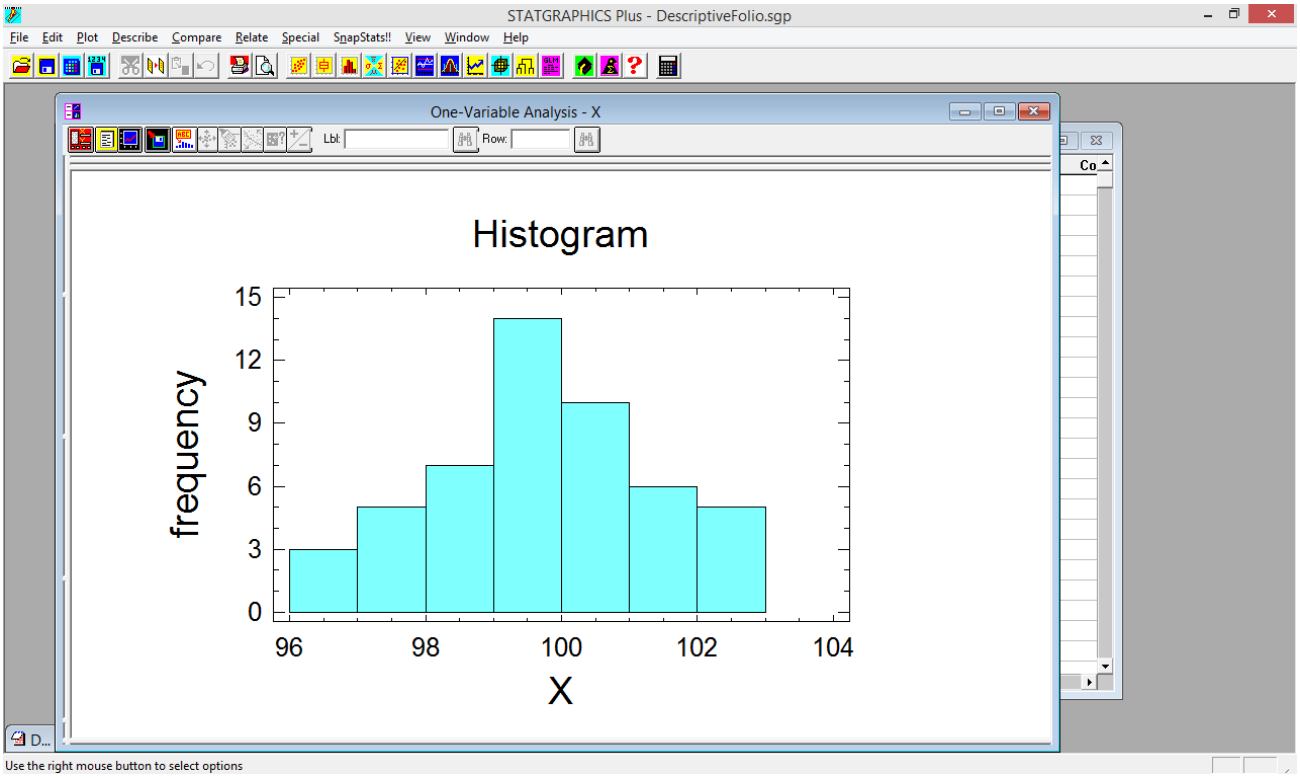


Рис. 14. Гістограма частот **Frequency Histogram**

Для зміни параметрів або виду графічного зображення, установите покажчик миші у вікні гістограми й клацніть правою клавішею миші. У діалоговому меню, що з'явився, виберете **пункт Pane** Options. З'явиться вікно **Frequency Plot Options** (рис. 15).

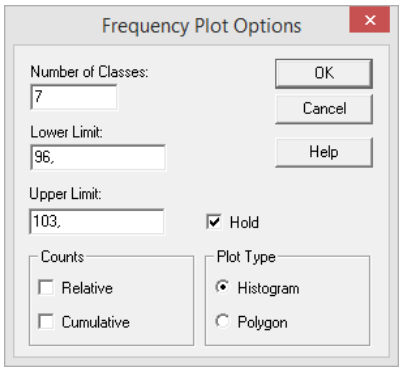


Рис. 15. Вікно зміни параметрів гістограми частот

Заповнивши відповідні поля можна змінити:

- число інтервалів розбивки (**Number of Classes**);
- нижню границю першого інтервалу розбивки (**Lower Limit**);
- верхню границю останнього інтервалу розбивки (**Upper Limit**).

Також можна змінити гістограму частот на:

- гістограму відносних частот (**Relative**);
- гістограму накопичених відносних частот (**Cumulative**).

Можлива зміна виду графічного зображення:

- гістограма (**Histogram**);
- полігон (**Polygon**).

Відзначивши потрібні опції, натисніть клавішу **OK**.

8. Складання звіту по роботі.

Система STATGRAPHICS *Plus* містить спеціальний засіб складання звіту по проведеній статистичній роботі **Statreporter**, що дозволяє здійснити комбінування тексту й графіки.

Для збереження результатів статистичної обробки даних в **Statreporter** необхідно клацнути правою кнопкою миші в будь-якому місці обраної панелі аналізу й в, що з'явився контекстному меню вибрати пункт **Copy Analysis to Statreporter…**

9. Збереження статистичного проекту виконаного завдання.

Для того щоб повторити всю схему й етапи проведеного статистичного аналізу для нового масиву даних система STATGRAPHICS *Plus* дозволяє зберігати результати роботи у вигляді спеціальних файлів статистичних проектів **Statfolio**. Для цього проводяться стандартні дії:

1. У головному меню виберіть команду **File | Save Statfolio As**;
2. У діалоговому вікні **Save Statfolio As** укажіть папку, у якій необхідно зберегти проект і ім'я файлу;
3. Натисніть кнопку **Save (Зберегти)**.

Для проведення статистичного аналізу за схемою збереженого проекту **Statfolio** необхідно відкрити його командою **File | Open Statfolio**, а потім завантажити новий файл даних, виконавши команду **File | Open Datafile**. Усі задані статистичні процедури будуть автоматично перелічені з новими даними.

III. Порядок проведення заключної частини заняття

Для того щоб завершити роботу в системі STATGRAPHICS *Plus* можна виконати одне з наступних дій:

1. Вибрати команду **File | Exit Statgraphics**;
2. Натиснути комбінацію клавіш <**Alt+F4**>;
3. Натиснути на кнопці закриття, яка перебуває в рядку заголовка вікна STATGRAPHICS *Plus*.