

Міністерство освіти і науки України  
Харківська державна академія фізичної культури  
Кафедра інформатики та біомеханіки

**В.С. Ашанін, С.С. Пятисоцька, Я.В. Жерновнікова, Ю.І. Петренко**

# **Системно-інформаційні основи наукових досліджень у фізичній культурі та спорті**

**Методичні рекомендації**

для здобувачів другого магістерського рівня вищої освіти  
спеціальності 017 «Фізична культура і спорт»

Харків  
ХДАФК  
2024

*Затверджено Вченою радою ХДАФК  
(Протокол № 7 від 25 червня 2024 року)*

**Рецензенти:** **О. О. Пилипко** – Завідувач кафедри водних видів спорту Харківської державної академії фізичної культури, кандидат педагогічних наук, професор;  
**А. М. Єфременко** – Завідувач кафедри легкої атлетики Харківської державної академії фізичної культури, кандидат наук з фізичного виховання та спорту, доцент.

### **Ашанін В. С.**

А 98 Системно-інформаційні основи наукових досліджень у фізичній культурі та спорті : методичні рекомендації для здобувачів другого магістерського рівня вищої освіти спеціальності 017 «Фізична культура і спорт» / Ашанін В. С., Пятисоцька С. С., Жерновнікова Я. В., Петренко Ю. І. – Харків: ХДАФК, 2024. – 144 с.

Методичні рекомендації призначені для здобувачів другого магістерського рівня вищої освіти спеціальності 017 «Фізична культура і спорт». Методичні рекомендації містять практичні завдання з дисципліни «Системно-інформаційні основи наукових досліджень у фізичній культурі та спорті» з використанням табличного процесора Microsoft Excel та системи «STATISTICA». Методичні рекомендації містять теоретичний матеріал (за необхідності), завдання до практичних робіт, а також послідовність їх виконання. Виконання практичних робіт дозволяє здобувачам вищої освіти закріпити теоретичний матеріал і набути практичних навиків автоматизації розрахунку методів математичної статистики.

УДК [796.011.2:796.072.22/004.771](07)

<b>Вступ</b> .....	4
<b>Практична робота №1.</b> Створення таблиць даних у програмах MS Excel та Statistica.....	6
<b>Практична робота №2.</b> Представлення статистичних даних у вигляді графіків та діаграм у програмі MS Excel.....	14
<b>Практична робота №3.</b> Побудова діаграм різного типу в програмі Statistica.....	19
<b>Практична робота №4.</b> Дескриптивна статистика у програмах MS Excel та Statistica.....	26
<b>Практична робота №5.</b> Параметричні методи перевірки статистичних гіпотез у програмах MS Excel та Statistica.....	37
<b>Практична робота №6.</b> Непараметричні методи перевірки статистичних гіпотез в програмі Statistica.....	43
<b>Практична робота №7.</b> Кореляційний аналіз із використанням параметричних методів у програмах MS Excel та Statistica.....	50
<b>Практична робота №8.</b> Використання непараметричних методів в кореляційному аналізі	58
<b>Практична робота №9.</b> Дисперсійний аналіз у програмах MS Excel та Statistica.....	61
<b>Практична робота №10.</b> Регресійний аналіз у програмі MS Excel.....	68
<b>Практична робота №11.</b> Регресійний аналіз у програмі Statistica.....	80
<b>Практична робота №12.</b> Факторний аналіз у програмі Statistica.....	89
<b>Практична робота №13.</b> Кластерний аналіз у програмі Statistica.....	104
<b>Практична робота №14.</b> Дискримінантний аналіз у програмі Statistica.....	116
<b>Список рекомендованої літератури</b> .....	141

## Вступ

Навчальна дисципліна «Системно-інформаційні основи наукових досліджень у фізичній культурі та спорті» є обов'язковою дисципліною освітньо-професійної програми підготовки здобувачів другого магістерського рівня вищої освіти спеціальності 017 «Фізична культура і спорт». Зміст дисципліни забезпечує спеціальну підготовку майбутніх тренерів відповідно сучасних вимог.

**Мета вивчення дисципліни:** формування наукового світогляду в процесі ознайомлення з положеннями теорії систем, розвиток системного мислення, забезпечення майбутніх фахівців знаннями та вміннями постановки системних досліджень, побудови моделей і застосування системно-інформаційного аналізу для розв'язання задач в галузі фізичної культури та спорту.

### Основні завдання навчальної дисципліни:

- ознайомити здобувачів освіти з теоретичною базою загальної теорії систем, теорією інформації, структурою та видами систем, закономірностями їх функціонування та розвитку;
- надати уявлення про основні етапи та методи системного аналізу;
- навчити здобувачів освіти загальним методам інформаційного моделювання об'єктів різноманітної природи;
- ознайомити здобувачів освіти із методами багатомірного аналізу даних;
- навчити сприйняттю інформації як системотворного феномену при вирішенні завдань професійної діяльності.

В результаті вивчення дисципліни здобувач освіти повинен **вміти**:

- проводити аналітичну обробку даних експериментальних досліджень в галузі фізичної культури та спорту з використанням статистичних програм;
- представити експериментальні дані та результати її обробки у графічному вигляді;
- формулювати статистично вірогідні висновки та готувати звітну документацію з комп'ютерного аналізу експериментальних даних;
- використовувати методи та моделі системно-інформаційного підходу при дослідженні властивостей людини, складних явищ, закономірностей управління та прийняття рішень в галузі фізичної культури та спорту в умовах недоліку інформації.

Відповідно до робочої програми навчальної дисципліни на практичні заняття відводиться 40 академічних годин, які розподілені на 14 практичних робіт.

Метою виконання практичних робіт є закріплення вивченого під час лекційних та практичних занять теоретичного матеріалу, набуття навичок самостійної математико-статистичної обробки емпіричних даних з використанням електронних таблиць Microsoft Excel та системи «STATISTICA», демонстрація зв'язку набутих знань з практичними ситуаціями досліджень в галузі фізичної культури та спорту.

## Практична робота № 1

### Тема «Створення таблиць даних у програмах MS Excel та Statistica»

**Мета заняття:** навчитись створювати таблиці даних у програмах MS Excel та Statistica.

#### Зміст і хід заняття

1. Практична робота з створення таблиць даних в програмі MS Excel.
2. Практична робота з створення таблиць даних в програмі Statistica. Виконання та захист практичного завдання.

### 1. Практична робота з створення таблиць даних в програмі MS Excel

#### Приклад 1.

1. Запустіть програму **MS Excel** знайшовши піктограму програми на робочому столі або у меню «Пуск» і клацніть на неї для запуску.

2. Огляд основних елементів інтерфейсу. **Меню** розташоване у верхній частині вікна програми і містить різноманітний набір команд. **Панель Інструментів** знаходиться під меню та містить кнопки швидкого доступу до часто використовуваних команд.

Документ Excel має назву «Робоча книга». Вона складається з окремих робочих аркушів, представлених у вигляді вкладок внизу вікна програми. Перемикання між аркушами відбувається шляхом натискання на відповідні вкладки. Кожен робочий аркуш являє собою електронну таблицю з 256 стовпцями та 65536 рядками. Стовпці позначаються латинськими літерами від А до Z, а рядки – цифрами. Кожна окрема комірka таблиці має унікальну адресу у вигляді поєднання літери стовпця та номера рядка. Наприклад, комірka на перетині другого стовпця (К) та другого рядка (6) матиме адресу К6. У кожен конкретний момент активованою є лише одна комірka таблиці, виділена рамкою. Активувати потрібну комірku можна шляхом кліку на ній мишею.

3. Створіть на **Аркуші1** електронну таблицю антропометричних показників спортсменок 15 років, які займаються різними видами стрибків за зразком (рис.1):

	A	B	C	D	E
1	<b>Вид стрибка</b>	<b>Зріст, см</b>	<b>Вага, кг</b>	<b>Довжина ніг, см</b>	<b>Розмах рук, см</b>
2	Стрибки у висоту	178	60	95	185
3	Стрибки у висоту	182	62	98	190
4	Стрибки у висоту	175	58	92	182
5	Стрибки у висоту	180	63	96	188
6	Стрибки у висоту	177	59	94	184
7	Стрибки з жердиною	188	68	102	195
8	Стрибки з жердиною	185	65	99	192
9	Стрибки з жердиною	190	70	104	198
10	Стрибки з жердиною	183	63	97	190
11	Стрибки з жердиною	187	67	101	194
12	Стрибки у довжину	175	62	92	180
13	Стрибки у довжину	178	64	94	183
14	Стрибки у довжину	172	60	90	177
15	Стрибки у довжину	180	66	96	185
16	Стрибки у довжину	174	61	91	179
17	Потрійний стрибок	176	63	93	181
18	Потрійний стрибок	179	65	95	184
19	Потрійний стрибок	173	61	91	178
20	Потрійний стрибок	181	67	97	186
21	Потрійний стрибок	175	62	92	180
22	Стрибки у висоту	179	61	96	186
23	Стрибки з жердиною	186	66	100	193
24	Стрибки у довжину	177	63	93	182
25	Потрійний стрибок	178	64	94	183
26	Стрибки у висоту	176	59	93	183
27	Стрибки з жердиною	189	69	103	196
28	Стрибки у довжину	174	61	91	179
29	Потрійний стрибок	180	66	96	185
30	Стрибки у висоту	177	60	94	184
31	Стрибки з жердиною	184	64	98	191

*Рис. 1. Таблиця з вхідними даними у MS Excel*

### **Основні правила створення статистичних таблиць:**

1. Таблиця повинна мати назву, яка коротко і точно відображає її зміст. Необхідно уникати великої кількості точок і ком у назві таблиці та граф, які ускладнюють читання. Якщо назва таблиці складається із двох і більше речень, точка ставиться із єдиною метою відділення речень один від одного, але не після останнього.

2. У кожній графі має бути зазначено її назву і одиниці виміру показників.

3. Якщо дані в таблиці мають дробові значення, то кількість знаків після коми має бути однаковою. Щоб застосувати формат комірки, виділіть потрібні комірки, перейдіть на вкладку **Основне** та виберіть потрібний формат у розділі **Число**. Або можливо клацнути правою кнопкою миші на комірці та вибрати **Формат комірок**.

4. Дані краще розташовувати в порядку зростання або за хронологією.

5. Заголовки рядків і граф повинні бути короткими і зрозумілими.

6. Для компактності доцільно використовувати прийоми укрупнення даних.

7. Бажано уникати порожніх граф у таблиці.

8. Використовуйте границі та сітку для виділення рядків та стовпців, щоб поліпшити читабельність таблиці.

9. Перед публікацією або використанням таблиці перевірте її на помилки та коректність даних.

У статистиці існують різні типи даних, які класифікуються на основі їх властивостей і можливостей аналізу.

1. *Кількісні дані (Quantitative Data)* – ці дані виражаються числовими значеннями і можуть бути виміряні. Вони поділяються на дві підкатегорії:

– дискретні дані (Discrete Data) – дані, які приймають окремі, чітко визначені значення (наприклад: кількість гравців у команді; кількість очок, набраних баскетбольною командою за гру; кількість перемог команди за сезон; кількість штрафних хвилин, отриманих гравцем у хокеї; кількість медалей, виграних спортсменом на змаганнях і т.д.).

– неперервні дані (Continuous Data) – дані, які можуть приймати будь-яке значення в певному інтервалі (наприклад: час, витрачений на подолання дистанції; вага спортсмена; відстань стрибка у довжину; температура тіла спортсмена під час тренування; пульс спортсмена під час бігу).

2. *Якісні дані (Qualitative Data)* – ці дані виражаються категоріями або описами, а не числами. Вони поділяються на дві підкатегорії:

– номінальні дані (Nominal Data) – дані, які представляють категорії без природного порядку або ранжування (наприклад: стать (чоловік/жінка); види спорту (футбол, баскетбол, теніс); позиції гравців у футбольній команді (воротар, захисник, півзахисник); назви команд («Динамо», «Шахтар», «Реал Мадрид»); типи змагань (чемпіонат, кубок, олімпіада)).

– порядкові дані (Ordinal Data) – дані, які представляють категорії з природним порядком або ранжуванням, але різниця між значеннями не має конкретного значення (наприклад: місця на змаганнях (1-е місце, 2-е місце, 3-е місце); ранги в турнірній таблиці (команда посіла 1-е місце в турнірній таблиці, інша команда – 2-е місце і т.д.); рівень підготовки спортсменів (початківець, середній рівень, професіонал); рейтинги суддів у змаганнях (оцінка виконання вправи: погано, задовільно, добре, відмінно); стадії змагань (кваліфікаційний раунд, чвертьфінал, півфінал, фінал)).

3. *Інтервальні дані (Interval Data)* – ці дані мають природний порядок і однаковий інтервал між значеннями, але не мають природного нуля. Тобто нуль не означає відсутність вимірюваної властивості (приклад: час початку матчів; рейтинг гравця в комп'ютерній грі або спортивній лізі за шкалою від 1 до 100; температура тіла спортсменів).



4. *Відносні дані (Ratio Data)* – ці дані мають природний порядок, однаковий інтервал між значеннями і природний нуль (наприклад, нуль кілограмів означає відсутність ваги), який означає відсутність вимірюваної властивості (наприклад: вага спортсменів; дистанція пробігу; час забігу та ін.).

4. Розуміння типів статистичних даних допомагає обрати правильні методи аналізу і візуалізації, а також дозволяє коректно інтерпретувати результати.

**Для введення даних** клацніть на клітинку, у яку ви хочете ввести дані. Введіть свої дані на клавіатурі. Натисніть клавішу **Enter**, щоб перейти на новий рядок, або **Tab**, щоб перейти на новий стовпець.

**Для редагування даних** двічі клацніть на клітинку з даними, яку ви хочете редагувати. Відредагуйте дані за необхідністю. Натисніть **Enter** для підтвердження змін або натисніть на іншу клітинку, щоб завершити редагування.

**Для надання назв змінним** виберіть верхній рядок у вашій таблиці, клацнувши на номері рядка або просто перетягнувши мишею по верхньому рядку. Клацніть правою кнопкою миші на обраному рядку та виберіть **Вставити**, щоб вставити новий рядок. Введіть назву стовпця у вставленому рядку, що тепер з'явився.

**Для надання границь таблиці** виділіть область, до якої потрібно додати границі клацнувши на комірці у верхньому лівому куті області, яку ви хочете включити в таблицю, та утримуйте кнопку миші, протягуючи до нижньої правої комірки.

5. Відкрийте вкладку **Головна**. У верхній частині програми Excel знаходиться рядок меню. Клацніть на вкладці **Головна**, щоб перейти на цю вкладку.

Оберіть опцію **Границі**. У групі інструментів **Шрифт**, яка зазвичай розташована в верхній частині вкладки **Головна**, знаходиться кнопка із зображенням рамки. Клацніть на стрілці поруч із цією кнопкою, щоб відкрити список опцій границь. Щоб вибрати тип границь у розкритому списку виберіть тип границь, які бажаєте застосувати до таблиці. Опції можуть включати звичайні горизонтальні та вертикальні лінії, зовнішні рамки та інші. Застосуйте границі після вибору бажаних опцій границь, клацніть на них. Границі будуть застосовані до виділеної області таблиці.

Збережіть зміни натиснувши на кнопку **Зберегти**.

6. Завдяки функції фільтрації даних у Excel можна легко знаходити потрібні дані та сортувати їх за певними критеріями, що значно спрощує

роботу з великими наборами даних. Щоб вставити фільтрацію даних у таблицю в Excel, необхідно виконати наступні кроки:

1. Відкрийте файл Excel, який містить таблицю даних.
2. Виділіть всю таблицю даних, включаючи заголовки стовпців, клацнувши та протягнувши курсор миші по всьому діапазону комірок.
3. Перейдіть на вкладку **Дані** на стрічці Excel.
4. У групі **Сортування та фільтр** знайдіть кнопку **Фільтр** та клацніть на неї.
5. Програма Excel автоматично додасть маленькі стрілочки вниз у кожен заголовок стовпця в таблиці. Це і є фільтри.
6. Щоб використати фільтрацію, клацніть на стрілочку вниз у заголовку стовпця, за яким ви хочете відфільтрувати дані.
7. З'явиться список усіх унікальних значень у цьому стовпці. Можна вибрати певні значення, натиснувши на них, щоб показати лише рядки з цими значеннями. Або ж можна скористатися опціями, такими як **Текстові фільтри** або **Числові фільтри**, щоб застосувати більш складні критерії фільтрації.
8. Після вибору значень для фільтрації натисніть ОК, і таблиця буде відфільтрована відповідно до обраних критеріїв.
9. Якщо необхідно видалити фільтр, то знову клацніть на стрілочку біля заголовка стовпця та оберіть опцію **Очистити фільтр**.
10. Можливо застосовувати фільтри до кількох стовпців одночасно, повторюючи кроки 6-8 для інших потрібних стовпців.
11. Щоб видалити всі фільтри з таблиці, перейдіть знову на вкладку **Дані**, у групі **Сортування та фільтр** натисніть кнопку **Очистити**.

## 2. Практична робота з створення таблиць даних в програмі Statistica

### Приклад 2.

1. Запустіть програму **Statistica** знайшовши піктограму програми на робочому столі або у меню **Пуск/Start menu** і клацніть на неї для запуску. У меню **Файл/File** виберіть **Створити/Create** або натисніть кнопку **Створити/Create** на панелі інструментів. У вікні **Створити новий документ/Create New Document** (рис. 2) виберіть вкладку **Таблиця/Spreadsheet**, де вкажіть необхідну кількість *змінних/variables* і *спостережень/cases* та натисніть кнопку **ОК**.

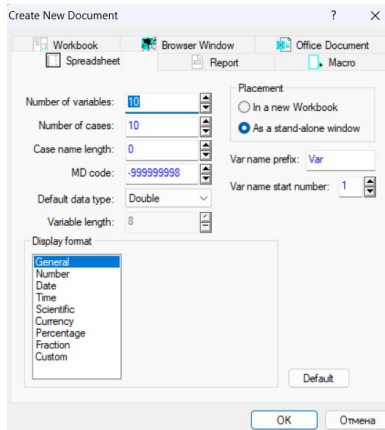


Рис. 2. Вікно команди Створити новий документ/Create New Document

2. Для введення даних у вікні **Таблиця/Spreadsheet** введіть дані в комірці. Для переміщення між комірками можна використовувати клавіші зі стрілками або клацнути мишкою на потрібній комірці. Для вставки нового рядка або стовпця клацніть правою кнопкою миші на номері рядка або стовпця та виберіть **Вставити/Insert**. Для видалення рядка або стовпця клацніть правою кнопкою миші на номері рядка або стовпця та виберіть **Видалити/Delete**.

3. На основі даних про спортивні заняття для дітей та молоді створіть таблицю за зразком (рис. 3):

	1	2	3	4	5	6
	Вид спорту	Категорія	Вік	Рівень	Кількість учасників	Приналежність
1	Футбол	Командний	12-16	Професійний	22	Клуб
2	Баскетбол	Командний	14-18	Аматорський	10	Школа
3	Теніс	Індивідуальний	10-14	Професійний	8	Клуб
4	Плавання	Індивідуальний	8-12	Аматорський	15	Секція
5	Гимнастика	Індивідуальний	6-10	Професійний	12	Клуб
6	Хокей	Командний	16-20	Професійний	20	Клуб
7	Волейбол	Командний	14-18	Аматорський	12	Школа
8	Бокс	Індивідуальний	16-20	Професійний	10	Клуб
9	Легка атлетика	Індивідуальний	12-16	Аматорський	25	Секція
10	Танці	Індивідуальний	8-12	Аматорський	18	Студія
11	Фігурне катання	Індивідуальний	10-14	Професійний	6	Клуб
12	Шахи	Індивідуальний	10-14	Аматорський	12	Секція
13	Карате	Індивідуальний	12-16	Аматорський	18	Секція
14	Велоспорт	Індивідуальний	14-18	Професійний	10	Клуб
15	Фехтування	Індивідуальний	14-18	Професійний	8	Клуб
16	Художня гімнастика	Індивідуальний	8-12	Професійний	10	Клуб
17	Регбі	Командний	16-20	Аматорський	15	Секція
18	Скелелазіння	Індивідуальний	12-16	Аматорський	8	Секція
19	Дзюдо	Індивідуальний	14-18	Аматорський	12	Секція
20	Настільний теніс	Індивідуальний	10-14	Аматорський	16	Клуб

Рис. 3. Таблиця з вхідними даними у Statistica

4. Для форматування таблиці виділіть один або кілька стовпчиків, клацнувши на їх заголовках. Для зміни ширини стовпчика перетягніть його праву межу. Для вирівнювання тексту в комірках використовуйте кнопки на панелі інструментів **Форматування**. Для об'єднання комірок виділіть їх та натисніть кнопку **Об'єднати** на панелі інструментів **Форматування**.

5. Для збереження таблиці у меню **Файл** виберіть **Зберегти**.

У вікні **Зберегти як/Save As** введіть ім'я файлу і місце для збереження та натисніть кнопку **Зберегти/Save**.

У програмі **Statistica** є кілька способів зберегти таблицю даних.

1. Зберегти як файл **Statistica Workbook (.stw)**:

– виберіть в меню **Файл > Зберегти як/File > Save As**;

– у вікні **Зберегти таблицю як/Save Spreadsheet As** виберіть тип файлу **Statistica Workbook Files (\*.stw)**;

– вкажіть ім'я файлу та місце збереження і натисніть **Зберегти/Save**.

2. Зберегти як **Excel файл (.xls або .xlsx)**:

– виберіть в меню **Файл > Експорт у/File > Export to**;

– у вікні експорту виберіть **Microsoft Excel files (\*.xls, \*.xlsx)**;

– вкажіть ім'я файлу та місце і натисніть **Експортувати/Export**.

3. Зберегти як **текстовий файл (.txt, .csv)**:

– у меню **Файл > Експорт у/File > Export to**;

– виберіть **Text files (\*.txt, \*.csv)**;

– налаштуйте опції експорту (роздільники, кодування тощо);

– вкажіть ім'я та місце і натисніть **Експортувати/Export**.

4. Зберегти як **HTML файл (.html)**:

– меню **Файл > Експорт у > HTML files (\*.html)**;

– налаштуйте опції експорту;

– вкажіть ім'я та місце, натисніть **Експортувати/Export**.

Файл **Statistica Workbook** зберігає всю інформацію про таблицю та зв'язки між даними. Решта форматів (**Excel**, **текст**, **HTML**) зберігають лише дані таблиці.

**Додаткові можливості:**

– додавання до таблиці формул, діаграм та інших об'єктів;

– імпорт даних з інших програм або з файлів;

– експорт даних в інші програми або файли.

6. Для того, щоб вставити фільтр для таблиці даних, виконайте наступні кроки:

1. Відкрийте потрібну таблицю даних у вікні **Spreadsheet**.

2. Для швидкого налаштування фільтру на панелі інструментів клацніть на вкладку **Дані/Data**. Виділіть всі стовпці або лише необхідні

для фільтрації та натисніть на кнопку **Автофільтр/Autofilter**. Цей тип авто фільтру створиться автоматично.

3. У вікні **Фільтр для/Filter for** виберіть діапазон стовпців або рядків, до яких ви хочете застосувати фільтр. Можете вибрати всю таблицю або лише певні стовпці/рядки.

4. Натисніть кнопку **Далі/Next**.

5. У наступному вікні виберіть умови фільтрування для кожного стовпця/рядка. Ви можете вказати числові діапазони, текстові рядки, дати тощо.

6. Після вибору всіх необхідних умов натисніть кнопку **ОК**.

Щоб видалити фільтр, знову клацніть на кнопку **Фільтр/Filter** на панелі інструментів і виберіть опцію **Зняти фільтр/Remove Filter**.

### Практична частина

1. Створіть таблицю даних у програмі MS Excel та у програмі Statistica за зразком. Додайте фільтрування для таких стовпців таблиці як: **Стать**, **Вік** та **Вид**.

Стать	Вік	Вид	Довжина тіла, см	Маса тіла, кг	Довжина рук, см	Розмах рук, см
Чоловік	20	Штовхання ядра	190	115	85	205
Чоловік	21	Штовхання ядра	188	110	82	200
Жінка	19	Штовхання ядра	178	92	75	185
Жінка	20	Штовхання ядра	182	95	78	190
Чоловік	22	Метання диска	192	105	88	210
Чоловік	21	Метання диска	188	100	84	205
Жінка	20	Метання диска	180	85	76	188
Жінка	19	Метання диска	175	80	72	182
Чоловік	22	Метання молота	195	120	90	215
Чоловік	21	Метання молота	190	115	87	210
Жінка	20	Метання молота	185	95	80	195
Жінка	19	Метання молота	182	90	77	190
Чоловік	22	Метання списа	188	95	85	205
Чоловік	21	Метання списа	185	90	82	200
Жінка	20	Метання списа	178	80	74	185
Жінка	19	Метання списа	175	75	71	180
Чоловік	22	Штовхання ядра	192	118	87	208
Чоловік	21	Метання диска	190	102	86	207
Жінка	20	Метання молота	183	92	78	192
Жінка	19	Метання списа	177	78	73	183
Чоловік	22	Штовхання ядра	188	112	84	202
Чоловік	21	Метання диска	186	98	82	200
Жінка	20	Метання молота	180	88	75	187
Жінка	19	Метання списа	174	72	70	177
Чоловік	22	Метання молота	193	117	89	212

## Практична робота № 2

### Тема «Представлення статистичних даних у вигляді графіків та діаграм у програмі MS Excel»

**Мета заняття:** ознайомитися із способами представлення даних у вигляді графіків та діаграм, навчитися створювати, редагувати та форматувати графічні об'єкти у програмі MS Excel.

#### Зміст і хід заняття

1. Загальні відомості про типи графіків у програмі MS Excel.
2. Практична робота із створення графічних об'єктів в програмі MS Excel.

Виконання та захист практичного завдання.

#### 1. Загальні відомості про типи графіків у програмі MS Excel

У програмі Microsoft Excel є багато різних типів графіків, які допомагають візуалізувати дані.

1. Лінійний графік (Line Chart) – використовується для відображення тенденцій даних протягом певного періоду часу. Він з'єднує точки даних лініями, що полегшує виявлення тенденцій та змін.

2. Стовпчастий графік (Column Chart) – використовують для порівняння величин різних категорій. Кожна категорія представлена вертикальними стовпцями.

3. Гістограма (Bar Chart) – схожа на стовпчастий графік, але стовпці розташовані горизонтально. Вона також використовується для порівняння величин різних категорій.

4. Кругова діаграма (Pie Chart) – показує частки кожної категорії від загальної суми. Вона зручна для візуалізації пропорцій.

5. Діаграма з областями (Area Chart) – діаграма схожа на лінійний графік, але області під лініями заповнені кольором. Це допомагає показати загальну суму і внесок кожної частини до неї.

6. Точковий графік (Scatter Plot) – використовується для відображення взаємозв'язку між двома змінними. Точки на графіку показують окремі значення.

7. Графік з бульбашками (Bubble Chart) – графік є розширенням точкового графіка. Кожна точка представлена бульбашкою, розмір якої залежить від третьої змінної.

8. Біржовий графік (Stock Chart) – використовується для відображення даних про акції, таких як ціна відкриття, закриття, найвища та найнижча ціни.

9. Поверхневий графік (Surface Chart) – використовується для показу тривимірних даних. Він дозволяє візуалізувати взаємозв'язок між трьома змінними.

10. Кільцевий графік (Doughnut Chart) – графік схожий на кругову діаграму, але з центральною частиною, що дозволяє показувати кілька наборів даних у вигляді шарів.

11. Графік з областями 3D (3D Area Chart) – графік з областями в тривимірному просторі дозволяє візуалізувати взаємозв'язок між трьома змінними з заповненими областями.

12. Радіальна діаграма (Radar Chart) – використовується для відображення багатовимірних даних у вигляді павутинки. Кожна вісь представляє одну змінну.

## 2. Практична робота із створення графічних об'єктів в програмі MS Excel

**Приклад 1.** По кількості забитих голів футбольними командами за сезон побудуйте лінійний графік.

Рішення.

1. Запустіть програму **MS Excel** знайшовши піктограму програми на робочому столі або у меню **Пуск** і клацніть на неї для запуску.

2. Створіть на **Аркуші 1** електронну таблицю за зразком (рис. 4):

	A	B	C	D	E	F	G	H
1		2000	2004	2008	2012	2016	2020	2024
2	<b>Країна 1</b>	15	17	18	22	28	25	32
3	<b>Країна 2</b>	13	10	9	11	12	10	7
4	<b>Країна 3</b>	23	19	15	15	17	19	20
5	<b>Країна 4</b>	12	10	13	8	9	11	13

*Рис. 4. Таблиця з вхідними даними у MS Excel*

3. Виділіть діапазон даних, включаючи заголовки стовпців і всі значення даних. Наприклад, виділіть клітинки A1:H5.

4. Перейдіть до вкладки **Вставка** у верхньому меню. У групі **Графіки** знайдіть і натисніть кнопку **Графік**. Виберіть підтип лінійного графіка (наприклад, **Лінійний** або **Лінійний з маркерами**).

5. За допомогою вкладки **Макет** або **Конструктор** (залежно від версії Excel) введіть назву. Натисніть на заголовок графіка і введіть новий заголовок «*Кількість медалей завойованих на Олімпійських іграх*».

6. Натисніть на графік, потім перейдіть до вкладки **Макет** або **Конструктор** (залежно від версії Excel). Виберіть Підписи **осей** і додайте

підписи для горизонтальної осі – *Рік* та вертикальної осі – *Кількість медалей, шт.*

7. Якщо легенда не відображається автоматично, додайте її, натиснувши **Легенда** у вкладці **Макет** або **Конструктор**, вибравши розташування легенди (за умови, якщо показників 2 і більше).

8. Форматування графіка. Для зміни кольору ліній натисніть на лінію, яку хочете змінити. У контекстному меню або в панелі форматування змініть колір лінії.

9. Додавання маркерів: Натисніть на лінію, виберіть **Формат серії даних**, а потім налаштуйте **маркери** на лінії.

10. Ви можете змінити товщину ліній, тип лінії, додати сітки і інші елементи за допомогою інструментів форматування.

11. Перевірте правильність відображення даних на графіку. Збережіть робочу книгу з графіком, натиснувши **Файл – Зберегти як**, виберіть розташування та введіть назву файлу. Побудований лінійний графік зображено на рис. 5.

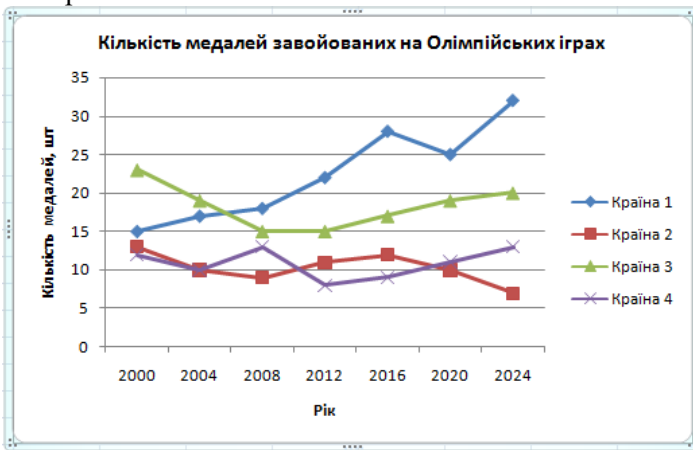


Рис. 5. Лінійний графік кількості медалей завойованих на Олімпійських іграх

**Приклад 2.** За показниками розподілу учасників міжобласних змагань по видам спорту побудуйте гістограму (стовпчасту діаграму).

Рішення.

1. Запустіть програму **MS Excel** знайшовши піктограму на робочому столі або у меню **Пуск** і клацніть на неї для запуску.

2. Створіть на **Аркуші 1** електронну таблицю за зразком (рис. 6):



	А	В
1	<b>Вид спорту</b>	<b>Кількість учасників</b>
2	Фехтування	20
3	Бокс	30
4	Гімнастика	40
5	Велоспорт	45
6	Теніс	55
7	Плавання	65
8	Волейбол	70
9	Баскетбол	85
10	Легка атлетика	95
11	Футбол	125

Рис. 6. Таблиця з вхідними даними у MS Excel

3. Виділіть діапазон даних, включаючи заголовки стовпців і всі значення даних. Наприклад, виділіть клітинки A1:B11.

4. Перейдіть до вкладки **Вставка** у верхньому меню. У групі **Графіки** знайдіть і натисніть кнопку **Гістограма**.

5. За допомогою вкладки **Макет** або **Конструктор** (залежно від версії Excel) введіть назву. Натисніть на заголовок графіка і введіть новий заголовок «Розподіл учасників міжобласних змагань по видам спорту».

6. Натисніть на графік, потім перейдіть до вкладки **Макет** або **Конструктор** (залежно від версії Excel). Виберіть **Підписи осей** і додайте підписи для горизонтальної осі – **Вид спорту**) та вертикальної осі – **Кількість учасників**).

7. Якщо легенда не відображається автоматично, додайте її, натиснувши **Легенда** у вкладці **Макет** або **Конструктор** і вибравши розташування легенди (за умови, якщо показників 2 і більше).

8. Зміна інтервалів. За необхідності натисніть правою кнопкою миші на горизонтальну вісь і виберіть **Формат осі**.

9. Ви можете змінити колір стовпчиків, додати сітки і інші елементи за допомогою інструментів форматування.

10. Натисніть на графік, потім перейдіть до вкладки **Макет** або **Конструктор** (залежно від версії Excel). Виберіть **Підписи даних**, вкажіть їх розміщення.

11. Перевірте правильність відображення даних на графіку. Збережіть робочу книгу з графіком, натиснувши **Файл – Зберегти як**, виберіть розташування та введіть назву файлу. Побудована гістограма (стовпчаста діаграма) зображена на рис. 7.

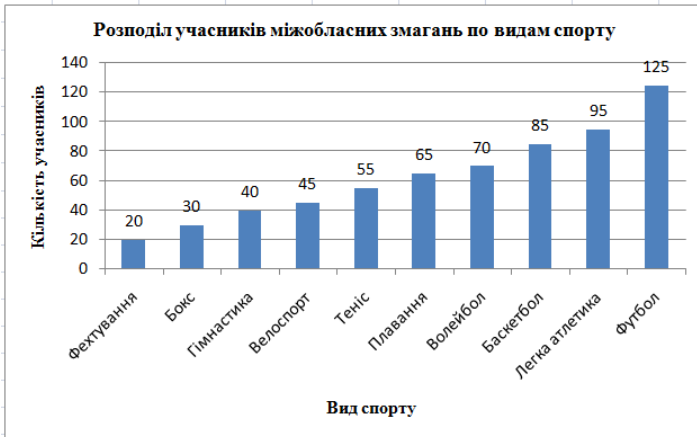


Рис. 7. Стопчастий графік розподілу учасників міжобласних змагань по видам спорту

### Практична частина

1. Побудуйте гістограму (лінійчасту діаграму) стрибків у висоту для різних спортсменів.

Спортсмен	Висота стрибка, см
Спортсмен 1	205
Спортсмен 2	202
Спортсмен 3	200
Спортсмен 4	198
Спортсмен 5	195

2. Побудуйте кругову діаграму за показниками розподілу уболівальників по видам спорту.

Вид спорту	Кількість уболівальників, млн.
Футбол	3,9
Баскетбол	1,2
Теніс	1,0
Бокс	0,8
Легка атлетика	0,6

3. Побудуйте точковий графік залежності маси тіла спортсмена та часу пробігу марафонської дистанції.

Спортсмен	Маса тіла, кг	Час пробігу, хв
Спортсмен 1	65	220
Спортсмен 2	68	230
Спортсмен 3	70	240
Спортсмен 4	75	250
Спортсмен 5	80	260

4. Побудуйте пелюсткову і бульбашкову діаграми за показником рухової активності футболістів протягом одного матчу. Проаналізуйте отримані дані.

Обсяг рухової активності (ОРА)			
Ігрове амплуа	хвилини	км	ОРА, %
Крайній захисник	70,9	8,7	20
Центральний захисник	69,0	9,0	21
Півзахисник	79,9	9,7	35
Крайній нападаючий	70,0	8,9	5
Центральний нападник	70,6	9,7	15

### Практична робота № 3

#### Тема «Побудова діаграм різного типу в програмі Statistica»

**Мета заняття:** ознайомитися із способами представлення даних у вигляді графіків та діаграм, навчитися створювати, редагувати та форматувати графічні об'єкти у програмі Statistica.

#### Зміст і хід заняття

1. Загальні відомості про типи графіків у програмі Statistica.
2. Практична робота із створення графічних об'єктів в програмі Statistica.

Виконання та захист практичного завдання.

#### 1. Загальні відомості про типи графіків у програмі Statistica

Програма Statistica пропонує широкий спектр типів графіків для візуалізації даних. Їх можна розділити на дві основні категорії.

##### 1. Користувацькі графіки:

- **2D Користувацькі графіки** – дозволяють створювати різні типи 2D-графіків, включаючи гістограми, лінійні графіки, графіки розсіювання та багато іншого.

- **3D Користувацькі графіки** – дозволяють створювати різні типи 3D-графіків, включаючи стовпчасті діаграми, лінійні графіки, графіки розсіювання та багато іншого.

- **3D Користувацькі діаграми розсіювання та поверхні** – дозволяють створювати 3D-діаграми розсіювання та графіки поверхонь.

- **Матричні графіки** – дозволяють створювати матричні графіки, які показують попарні відносини між декількома змінними.

- **Піктограми** – дозволяють створювати піктограми, які представляють категорійні дані за допомогою символів.

## 2. Статистичні (та швидкі статистичні) графіки:

- **Статистичні графіки** – автоматично генерується на основі вибраного вами статистичного аналізу. Наприклад, якщо ви виконаєте t-тест, Statistica автоматично створить t-тестову діаграму.

- **Швидкі статистичні графіки:** Цей тип графіка дозволяє швидко створювати прості графіки, такі як гістограми, лінійні графіки та графіки розсіювання.

Окрім цих двох основних категорій, Statistica також пропонує ряд спеціалізованих графіків, таких як графіки контролю якості, карти потоків процесів та геопросторові графіки.

### Вибір типу графіка

Тип вибраного графіка залежить від типу даних, та від того, що необхідно дослідити за його допомогою.

*Для категорійних даних* використовуються гістограми, стовпчасті діаграми або піктограми.

*Для кількісних даних* використовуються лінійні графіки, графіки розсіювання або коробчасті діаграми.

*Для порівняння груп даних* використовуються стовпчасті діаграми, лінійні графіки або графіки розсіювання.

*Для показу тенденцій з часом* використовуються лінійні графіки або ковзні середні.

*Для виявлення зв'язків між змінними* використовуються графіки розсіювання або кореляційні матриці.

Обидві групи графіків, і Статистичні, і Користувальницькі можуть бути викликані такими основними способами:

- за допомогою кнопок на панелі інструментів електронних таблиць;
- за допомогою меню (як контекстних, викликаних за допомогою правої кнопки миші, так і за допомогою основного меню *Графіки/Graphs*) електронних таблиць.

Графіки із програми Statistica зберігаються у файлах з розширенням \*.stg. Після того як графік побудований, відповідний графічний документ може бути збережений як:

- графічний документ у спеціальному графічному форматі програми Statistica (розширення файлу \*.stg), який може бути відкритий пізніше та використаний у процесі статистичного аналізу;
- графічний документ у растровому графічному форматі (формат \*.bmp, \*.pcx);
- графічний документ у графічному форматі метафайлу Windows (\*.wmf).

Графіки можуть бути виведені на принтер, слайди та інші носії з високою роздільною здатністю.

## 2. Практична робота із створення графічних об'єктів в програмі Statistica

**Приклад 1.** За показниками кількості аудиторії та доходу за видами спорту побудуйте стовпчикову діаграму.

Рішення.

1. Запустіть програму **Statistica** знайшовши піктограму програми на робочому столі або у меню **Пуск/Start menu** і клацніть на ній для запуску. У меню **Файл/File** виберіть **Створити/Create** або натисніть кнопку **Створити/Create** на панелі інструментів. У вікні **Створити новий документ/Create New Document** виберіть вкладку **Таблиця/Spreadsheet**, де вкажіть необхідну кількість **Змінних/Variables** і **Спостережень/Cases** та натисніть кнопку **ОК**.

2. На основі даних про кількість аудиторії та доходу за видами спорту створіть таблицю за зразком (рис. 8):

Вид спорту	1	2
	Кількість глядачів, тис	Прибуток, млн.
Футбол	85	120
Хокей	70	95
Бокс	65	90
Баскетбол	55	75
<b>Теніс</b>	40	<b>65</b>

Рис. 8. Таблиця з вхідними даними у програмі Statistica

3. Відкрийте стартове вікно модуля **Graphs/Графіки** у верхньому меню. Виберіть **2D - Bar Column Plots /Стовпчикові діаграми** (рис. 9).

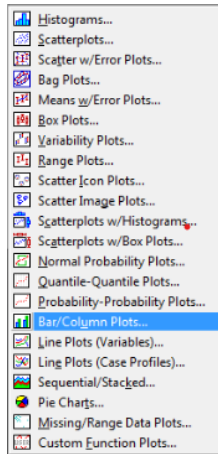


Рис. 9. Вікно модуля *Graphs/Графіки*, *2D-Bar Column Plots /Стовпчикові діаграми*

4. У вікні налаштувань діаграми виберіть **Variables/Змінні** для стовпців з даними, які необхідно відобразити, або **Виберіть всі/Selekt all** (рис. 10).

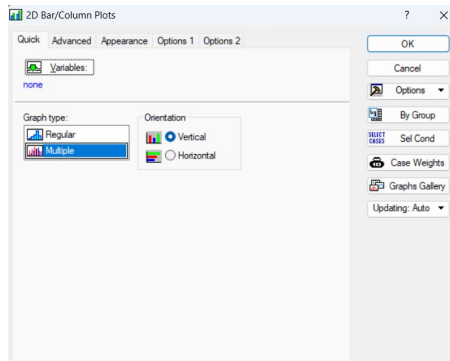


Рис. 10. Вікно налаштувань діаграми *2D-Bar Column Plots /Стовпчикові діаграми*

5. На вкладці **Тип графіка** виберіть **Складовий/Multiple**, орієнтацію **Вертикальну/Vertical**. Натисніть **ОК**. Клацнувши курсором мишки на створеній діаграмі відкриється вкладка **All graph properties/Всі параметри графіка** (рис. 11). Додаткові налаштування включають кольори стовпців, підписи осей, легенду тощо. Результат побудови стовпчикової діаграми показано на рисунку 12.

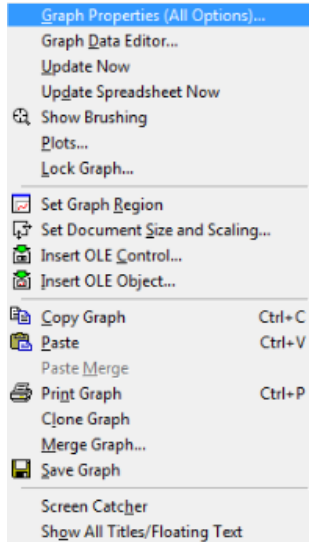


Рис. 11. Вкладка All graph properties/Всі параметри графіка

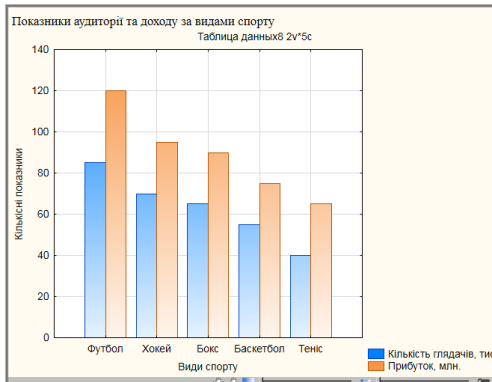


Рис. 12. Стовпчикова діаграма показників кількості аудиторії та доходу за видами спорту

6. Для того, щоб зберегти діаграму необхідно вибрати команду **File/Файл - Save As/Зберегти як** обравши відповідний формат (наприклад, .png, .jpg або інші). Для експорту у звіт або презентацію, можна скопіювати діаграму і вставити її у відповідний документ.

**Приклад 2.** Створіть лінійний графік за результатами плавання двох спортсменів на дистанції 50 метрів протягом декількох днів тренувань.

Рішення.

1. Запустіть програму **Statistica** знайшовши піктограму програми на робочому столі або у меню **Пуск/Start menu** і клацніть на ній для запуску. У меню **Файл/File** виберіть **Створити/Create** або натисніть кнопку **Створити/Create** на панелі інструментів. У вікні **Створити новий документ/Create New Document** виберіть вкладку **Таблиця/Spreadsheet**, де вкажіть необхідну кількість **Змінних/Variables** і **Спостережень/Cases** та натисніть кнопку **ОК**.

2. На основі даних результатів плавання двох спортсменів на дистанції 50 метрів протягом декількох днів тренувань заповніть електронну таблицю за зразком (рис. 13).

	1 Плавець 1, с	2 Плавець 2, с
1	28,5	29,2
2	28,1	28,9
3	27,8	29,1
4	27,5	28,3
5	27,2	28
6	26,9	28,1
7	26,6	27,4
8	26,3	27,1
9	26,0	26,8
10	25,7	26,5

Рис. 13. Таблиця з вихідними даними у програмі Statistica

3. Відкрийте стартове вікно модуля **Graphs/Графіки** у верхньому меню. Виберіть **2D – Line Graphs/Лінійні графіки**.

4. У вікні налаштувань діаграми виберіть **Variables/Змінні** для стовпців з даними, які необхідно відобразити, або **Виберіть всі/Selekt all**.

5. На вкладці **Тип графіка** виберіть **Складовий/Multiple**. Натисніть **ОК**. Результат побудови лінійного графіку показано на рисунку 14.

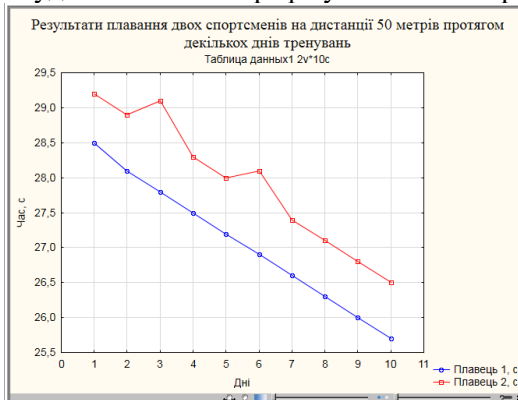


Рис. 14. Лінійний графік результатів плавання двох спортсменів на дистанції 50 метрів протягом декількох днів тренувань



6. Клацнувши курсором мишки на створеній діаграмі відкриється вкладка **All graph parameters/Всі параметри графіка**.

7. Встановіть кольори для кожної лінії, щоб їх було легко розрізнити. Додайте підписи до осей Дні на осі X і Час (с) на осі Y (клацнувши окремо на кожній вісі). Додайте легенду, яка пояснює, яка лінія відповідає якій дистанції. Виберіть необхідний стиль та товщину ліній.

8. Для того, щоб зберегти діаграму необхідно вибрати команду **File/Файл - Save As/Зберегти як** обравши відповідний формат (наприклад, .png, .jpg або інші). Для експорту у звіт або презентацію, можна скопіювати діаграму і вставити її у відповідний документ

### Практична частина

1. За показниками турнірної таблиці кращих гравців команди з футзалу побудуйте:

– 2D діаграму розсіювання за показниками гольових пасів та забитих голів;

– 2D діаграму розмаху за показниками зіграних ігор та забитих голів;

– 2D лінійну діаграму за показником кількості жовтих карток.

– 3D XYZ діаграму розсіювання за показниками зіграних ігор, забитих голів та гольових пасів.

– 3D XYZ графік поверхностей за показниками зіграних ігор, забитих голів та кількістю отриманих жовтих карток.

Гравець	Зіграні ігри	Забиті голи	Гольові паси	Жовта картка
Гравець 1	25	24	15	5
Гравець 2	30	25	15	3
Гравець 3	22	15	12	2
Гравець 4	24	16	14	5
Гравець 5	24	13	12	3
Гравець 6	22	22	10	5
Гравець 7	21	20	11	4

2. За допомогою діаграми розсіювання відобразіть взаємозв'язок між відстанню стрибка в довжину та часом пробігу на 100 метрів для 10 спортсменів.

Відстань стрибка, м	Час на 100 м, с
6,5	12,5
6,8	12,3
7,1	12,1
6,4	12,6
7	12,2
7,3	11,9
6,6	12,4
6,9	12

7,2	12,1
6,7	12,5

3. Побудуйте кругову діаграму на основі даних про кількість учнів, що відвідують заняття з різних видів спорту в Дитячо-юнацькій спортивній школі.

Види спорту	Кількість учнів
Футбол	84
Баскетбол	68
Волейбол	56
Хокей	32
Гандбол	25
Регбі	22

### Практична робота № 4

#### Тема «Дескриптивна статистика у програмах MS Excel та Statistica»

**Мета заняття:** навчитися розрахунку дескриптивної статистики в програмах MS Excel та Statistica.

#### Зміст і хід заняття

1. Практична робота із розрахунку показників дескриптивної статистики в програмі MS Excel.
2. Практична робота із розрахунку показників дескриптивної статистики в програмі Statistica.

Виконання та захист практичного завдання.

#### 1. Практична робота із розрахунку показників дескриптивної статистики в програмі MS Excel

Дескриптивна статистика (англ. descriptive statistics) або описова статистика – це розділ статистики, який вивчає методи збору, зведення та наочного представлення (у вигляді таблиць, графіків) емпіричних даних з метою їх подальшого аналізу.

Дослідження у галузі фізичної культури і спорту ґрунтуються на спостереженнях, експерименті та тестуванні. Значна частина наукових методів спирається на результати вимірювань великих груп спортсменів. Практика ФКС має у своєму розпорядженні вихідні дані у вигляді статистичної сукупності, де її поодинокі показники відображають досягнення конкретного спортсмена, а їх варіювання свідчить про індивідуальну відмінність спортсменів за вимірюваним показником.

Основним завданням дескриптивної статистики є узагальнення числових характеристик вибірки, таких як показники центральної тенденції (середнє арифметичне, медіана, мода), показники мінливості

(розмах варіації, дисперсія, середнє квадратичне відхилення та ін.) та показники форми розподілу даних (асиметрія, ексцес). Отримані статистичні характеристики дають уявлення про загальні властивості досліджуваної сукупності та використовуються для подальшого застосування методів аналізу даних та побудови статистичних моделей.

До основних показників дескриптивної статистики відносяться:

1. Середнє значення (MEAN) – сума всіх значень, поділена на їх кількість. Показує центральну тенденцію розподілу.

2. Медіана (MEDIAN) – значення посередині відсортованого ряду даних. Ділить розподіл навпіл.

3. Мода (MODE) – значення, яке зустрічається найчастіше в ряді даних. Показує найбільш типове значення.

4. Розмах варіації (RANGE) – різниця між найбільшим та найменшим значеннями ряду. Показує варіацію даних.

5. Дисперсія (VARIANCE) – середнє значення квадратів відхилень від середнього. Характеризує відхилення значень.

6. Стандартне відхилення (STD DEV) – квадратний корінь з дисперсії. Показує типове або очікуване відхилення від середнього.

7. Ексцес (KURTOSIS) – міра того, наскільки розподіл даних має «хвости» порівняно з нормальним розподілом.

8. Коефіцієнт асиметрії (SKEWNESS) – міра того, наскільки розподіл даних нахилений вправо або вліво.

9. Коефіцієнт варіації (COEFFICIENT OF VARIATION) – відношення середньоквадратичного відхилення до середнього арифметичного, виражене у відсотках. Чим більше значення коефіцієнта варіації, тим більший розкид даних навколо середнього.

Існують певні критерії для інтерпретації значень коефіцієнта варіації:

–  $CV < 10\%$  – низький рівень варіації: вказує на дуже однорідні дані з невеликим розкидом навколо середнього значення. Це може свідчити про високу точність та якість даних;

–  $10\% \leq CV < 20\%$  – помірний рівень варіації: вважається прийнятною для більшості випадків. Дані мають відносно невеликий розкид, але все ще достатньо однорідні;

–  $20\% \leq CV < 30\%$  – високий рівень варіації: вказує на значний розкид даних навколо середнього значення. Такий рівень варіації може бути прийнятним для певних видів даних, але в деяких випадках може вимагати подальшого дослідження причин високої мінливості.

–  $CV \geq 30\%$  – дуже високий рівень варіації: вказує на надзвичайно високу мінливість даних. Це може бути викликано наявністю аномальних

значень (викидів), недостовірністю даних або природною високою варіабельністю процесу, що вивчається.

Ці критерії є загальними рекомендаціями, але інтерпретація значень коефіцієнта варіації може залежати від конкретної галузі, типу даних та мети дослідження.

**Приклад 1.** Зібрано дані про 10 легкоатлетів - вік, зріст, вага, час бігу на 100 метрів. Використовуючи дану таблицю виконайте розрахунок показників дескриптивної статистики у програмі Excel. Використовуючи стовпці «Вік» і «Час бігу на 100 м», побудуйте точкову діаграму для візуалізації зв'язку між цими показниками та зробіть висновок.

Рішення.

Визначення показників дескриптивної (описової) статистики можливе кількома методами. За допомогою інструменту «**Пакет аналізу**», а також **Майстра функцій, категорія Статистичні**. Розглянемо перший варіант з використанням команди **Аналіз даних**.

1. Запустіть програму **MS Excel** знайшовши піктограму програми на робочому столі або у меню **Пуск** і клацніть на неї для запуску.

2. Створіть на **Аркуші1** електронну таблицю з вхідними даними за зразком (рис.15):

	A	B	C	D	E
1	<b>Спортсмени</b>	<b>Вік</b>	<b>Зріст, см</b>	<b>Вага, кг</b>	<b>Час бігу на 100 м, с</b>
2	Іван З.	22	182	75	10,8
3	Марія К.	25	170	60	11,5
4	Петро Н.	20	188	80	10,2
5	Камілла П.	24	175	65	11,1
6	Андрій О.	21	180	72	10,6
7	Наталія Д.	23	168	58	11,7
8	Максим С.	26	185	78	10,4
9	Катерина Г.	22	172	62	11,3
10	Сергій М.	24	183	75	10,5
11	Ольга В.	21	167	57	11,9

*Рис. 15. Таблиця з вхідними даними у MS Excel*

3. Знайдіть у групі команд **Дані** команду **Аналіз даних**. Натисніть кнопку **ОК**.

4. В **Інструментах аналізу** знайдіть вкладку **Описова статистика**. Натисніть кнопку **ОК**.

5. У вікно **Вхідний інтервал** внесіть коди створеної таблиці, виділивши цю таблицю (лише числові показники). Спосіб **Групування** вкажіть – по стовпцях.

6. В області **Параметри виводу** виберіть режим **Вихідний інтервал** – після цього в доступному вікні введіть (або вкажіть клацанням миші на

аркуші) адресу комірки, де повинен з'явитися результат чи відмітьте у параметрах виводу даних **Новий робочий лист**. Поставте позначку навпроти **Підсумкової статистики**. Натисніть кнопку **ОК**.

7. Отримані результати обробки даних – для наочності представлені на рисунку 16.

	A	B	C	D	E	F	G	H
1	Вік		Зріст, см		Вага, кг		Час бігу на 100 м, с	
2								
3	Среднее	22,8	Среднее	177	Среднее	68,2	Среднее	11
4	Стандартная ошибка	0,61101009	Стандартная оши	2,389793	Стандартная ош	2,764055	Стандартная ошибк	0,185592145
5	Медиана	22,5	Медиана	177,5	Медиана	68,5	Медиана	10,95
6	Мода	22	Мода	#Н/Д	Мода	75	Мода	#Н/Д
7	Стандартное отклонение	1,93218357	Стандартное отк	7,557189	Стандартное отк	8,7407094	Стандартное откло	0,586893895
8	Дисперсия выборки	3,73333333	Дисперсия выбо	57,11111	Дисперсия выбс	76,4	Дисперсия выборки	0,344444444
9	Эксцесс	-0,96141582	Эксцесс	-1,63982	Эксцесс	-1,880426	Эксцесс	-1,403649472
10	Асимметричность	0,23566971	Асимметричност	0,001931	Асимметричнос	-0,0042928	Асимметричность	0,206115182
11	Интервал	6	Интервал	21	Интервал	23	Интервал	1,7
12	Минимум	20	Минимум	167	Минимум	57	Минимум	10,2
13	Максимум	26	Максимум	188	Максимум	80	Максимум	11,9
14	Сумма	228	Сумма	1770	Сумма	682	Сумма	110
15	Счет	10	Счет	10	Счет	10	Счет	10
16	Уровень надежности(95,0%	1,38220085	Уровень надежн	5,406088	Уровень надежн	6,2527268	Уровень надежности	0,4198386

*Рис. 16. Результат роботи інструменту Описова статистика*

8. Пакет аналізу даних містить у собі розрахунок більшості показників дескриптивної статистики. Проте, значення коефіцієнту варіації необхідно розрахувати самостійно, використовуючи наступні значення (**Стандартне відхилення / Середнє арифметичне**) \* 100%.

9. Для розрахунку коефіцієнту варіації за показником **Вік**, клацніть на вільну комірку для вводу формули, з клавіатури введіть знак **Дорівнює** і поставте одну відкриту дужку. Клацніть на розраховане значення **Стандартного відхилення для Віку** у отриманій результативній таблиці описових статистик. Введіть знак ділення з клавіатури, клацніть на розраховане значення **Середнього арифметичного для Віку** у тому самому стовпчику. Введіть з клавіатури закриту дужку, знак множення та показник 100. Натисніть **Enter**. Отриманий показник оцінюється у відсотках – 8,47%, що вказує на низький рівень варіації. Таким чином, коефіцієнт варіації можна розрахувати для кожного показника, копіюючи формулу. Отримано наступні дані: зріст – 4,26%, вага – 12,8%, час бігу на 100 метрів – 5,3%.

10. Для візуалізації результатів проведеної дескриптивної статистики виділіть числові показники стовпчика **Вік**. Натисніть на вкладку **Вставити – Діаграми**. Виберіть діаграму під назвою **Ящик з вусами** (рис. 17), натисніть **ОК**.

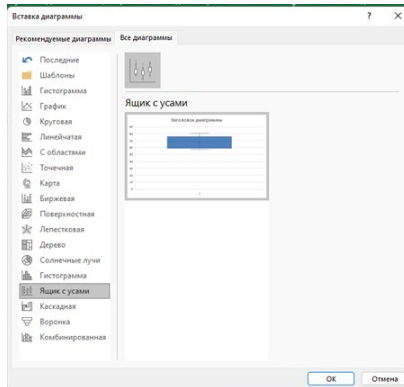


Рис.17. Вікно для вставки діаграми «Ящик з вусами»

Таким же чином побудуйте діаграму за показниками *Зріст*. Результати побудови діаграм зображено на рис 18.

Ця діаграма складається з прямокутника, що представляє проміжок між першим і третім квантилями (тобто міжквартильний розмах), а також «вусів» – відрізків, що зображують мінімальне і максимальне значення в наборі даних, виключаючи викиди. Вона компактно відображає розподіл, центральну тенденцію та розкид даних, що допомагає візуально порівняти кілька наборів даних. Ящик з вусами зручний для швидкого аналізу та виявлення можливих викидів.

На побудованих діаграмах можна побачити:

- межі прямокутника – перший і третій квантилі;
- риска всередині прямокутника – значення медіани;
- хрестик/коло – відображає середнє арифметичне значення.

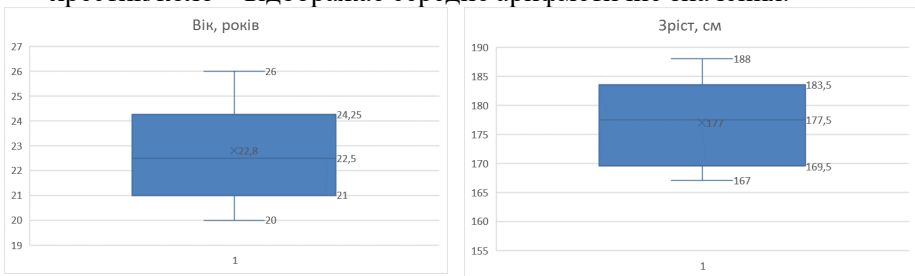
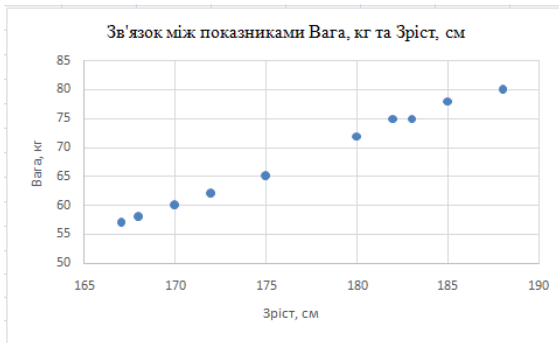


Рис. 18. Графічна інтерпретація результату побудови діаграм «Ящик з вусами» для показників *Вік* і *Зріст*

10. Побудуйте точкову діаграму використовуючи стовпці **Вага** і **Зріст** (рис. 19).



*Рис. 19. Графічна інтерпретація результату побудови точкової діаграми за показниками Вага і Зріст*

**Інтерпретація результатів.** Аналіз дескриптивної статистики показує, що група легкоатлетів є відносно однорідною. Середній вік легкоатлетів становить 22,8 років, що відповідає очікуваному віковому діапазону для спортсменів високого рівня. Середній зріст 177 см та середня вага 68,2 кг є типовими для даного виду спорту. Середній час бігу на 100 м – 11,0 секунд, що свідчить про високий рівень швидкісних якостей групи. Для більшості показників медіана та мода близькі до середнього значення, що свідчить про відносну симетричність розподілу даних.

Найшвидший час бігу на 100 м – 10,2 секунди, а найповільніший – 11,9 секунди, що відображає значну різницю в швидкісних якостях.

Вік, зріст та час бігу на 100 метрів мають низькі коефіцієнти варіації (8,2%, 4,26% та 5,3% відповідно), що свідчить про відносну однорідність групи за цими показниками. Вага має помірний коефіцієнт варіації (12,8%), що є прийнятним для групи спортсменів.

Візуальний аналіз графіку показує, що наявна чітка пряма лінійна залежність між вагою та зростом спортсменів. Точки на графіку утворюють виразну пряму лінію, що свідчить про сильний лінійний зв'язок.

Загалом, точковий графік підтверджує висновок про те, що чим вище спортсмен, тим він може мати більшу вагу. На результат можуть впливати інші чинники, такі як генетичні особливості, рівень тренувань та ін.

## **2. Практична робота із розрахунку показників дескриптивної статистики в програмі Statistica**

**Приклад 1.** За результатами даних спортсменок артистичного плавання (n=10) розрахуйте показники дескриптивної статистики та зробіть висновки.

Рішення.

1. Запустіть програму **Statistica** знайшовши піктограму програми на робочому столі або у меню **Пуск** і клацніть на неї для запуску. У меню **Файл** виберіть **Створити** або натисніть кнопку **Створити** на панелі інструментів.

2. У вікні **Створити новий документ** виберіть вкладку **Таблиця**, де вкажіть необхідну кількість **змінних** і **спостережень** та натисніть кнопку **ОК**. Заповніть електронну таблицю з вхідними даними за зразком (рис. 20).

	1 Вік	2 Зріст, см	3 Вага, кг	4 Бали за технічну складову	5 Бали за артистичну складову
Ганна Р.	18	168	52	87,5	89,2
Вікторія К.	20	172	55	92,3	91,7
Катерина Ш.	19	170	54	88,9	90,5
Марія Б.	21	175	58	94,1	93,8
Юлія В.	22	169	53	90,2	88,6
Олена Т.	20	171	56	91,8	92,1
Ірина А.	19	168	51	88,4	89,3
Наталія О.	21	174	57	93,5	92,9
Олександра М.	18	167	50	86,7	87,4
Валерія Ф.	20	173	56	91,6	90,8

Рис. 20. Таблиця з вхідними даними у Statistica

3. Відкрийте стартове вікно модуля **Basic Statistics/Tables – Основні статистики/Таблиці** (рис. 21).

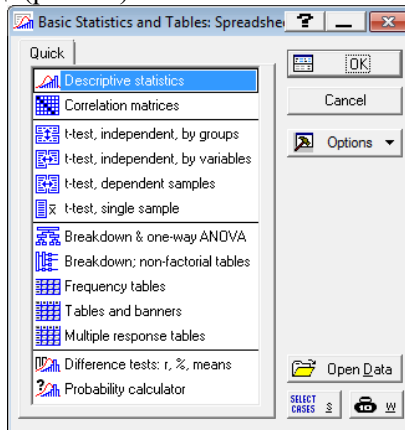


Рис. 21. Стартове вікно модулю *Основні статистики/Таблиці* системи Statistica



4. Клацніть у стартовому вікні модуля на рядку **Descriptive statistics/Описова статистика**. Наступне вікно з'явиться на екрані (рис. 22).

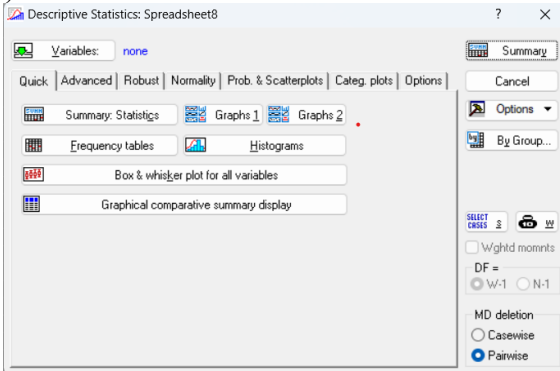


Рис. 22. Початкове вікно діалогу *Descriptive statistics/Описова статистика*

5. Натисканням кнопки **Variables/Змінні** розкрийте вікно вибору змінних. Виберіть всі три змінні. Далі клацніть **OK** або натисніть **Enter** на клавіатурі.

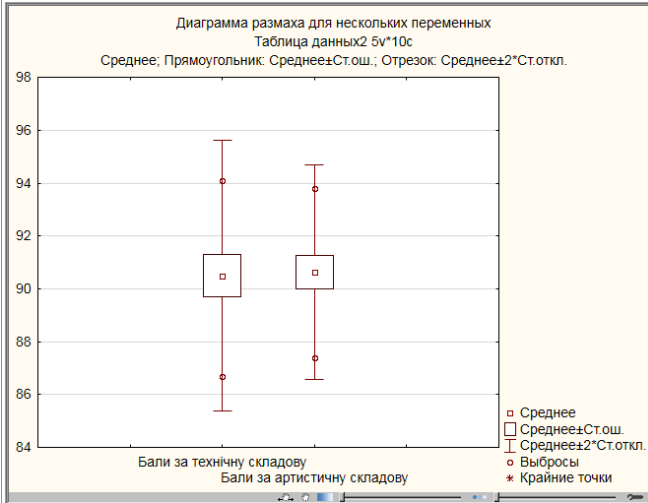
6. За замовчуванням програма розраховує середнє арифметичне значення, стандартне відхилення, а також визначаються кількість спостережень, максимальне і мінімальне значення. Якщо потрібно розрахувати інші показники описової статистики, то слід перейти на вкладку **Advanced/Додатково** і встановити позначку біля необхідних показників. Наприклад – медіана, мода, коефіцієнт варіації та ін. Розрахунок показників ініціюється натисканням кнопки **OK**, після чого з'явиться вікно з таблицею результатів на ім'я **Descriptive statistics/Описова статистика/Таблиця даних** (рис. 23).

Перемінная	Описательные статистики (Таблица данных2)									
	Среднее	Медиана	Мода	Частота моды	Минимум	Максим.	Ст. откл.	Козф. Вар	Асимметрия	Эксцесс
Вік	19,8000	20,0000	20,00000	3	18,0000	22,0000	1,316561	6,649299	0,087641	-0,75127
Зріст, см	170,7000	170,5000	168,0000	2	167,0000	175,0000	2,750757	1,611457	0,230613	-1,31402
Вага, кг	54,2000	54,5000	56,00000	2	50,0000	58,0000	2,658320	4,904650	-0,223577	-1,13608
Бали за технічну складову	90,5000	90,9000	Множест.	1	86,7000	94,1000	2,551688	2,819545	-0,102321	-1,35694
Бали за артистичну складову	90,6300	90,6500	Множест.	1	87,4000	93,8000	2,025422	2,234825	0,010294	-0,88113

Рис. 23. Підсумкова таблиця *Descriptive statistics/Описової статистики*

7. Після розрахунку вікно з результатами можна просто закрити, а можна зберегти у форматі *STW*. Надалі до збережених результатів можна буде звернутися у будь-який час, не переглядаючи при цьому первинні дані.

8. Побудуйте діаграму розмаху для показників **Бали за технічну складову** та **Бали за артистичну складову**, відмітьте опцію Середнє/SD і натисніть кнопку **ОК** (рис 24).



*Рис. 24. Графічна інтерпретація результату побудови діаграми розсіювання за показниками балів за технічну та артистичну складові у артистичному плаванні*

**Інтерпретація результатів.** За результатами аналізу описової статистики показників спортсменок з артистичного плавання визначено, що середній вік спортсменок – 19,8 років, це відповідає очікуваному віковому діапазону для цього виду спорту. Середній зріст 170,7 см та середня вага 54,2 кг є типовими для артистичного плавання. Середні бали за технічну (90,5) та артистичну (90,6) складові є досить високими, що вказує на високий рівень майстерності групи. Стандартні відхилення для більшості показників є невеликими, що свідчить про відносну однорідність групи.

Низькі значення коефіцієнтів варіації для всіх показників свідчать про невелику мінливість даних.

Для віку, зросту та ваги медіана та мода близькі до середнього, що вказує на симетричний розподіл. Для балів за технічну та артистичну складові медіана та мода дещо різняться, що може свідчити про незначну асиметрію.

Загалом, проведений аналіз дескриптивної статистики свідчить про те, що група спортсменок є відносно однорідною за віком, зростом, вагою та оцінками за виступи. Спостерігається незначна асиметрія в розподілі

даних для оцінок за виступи. Існують певні відмінності в оцінках між окремими спортсменками, що може бути пов'язано з різним рівнем підготовки та талантом. Вік має слабкий позитивний зв'язок з балами за технічну складову, але не є визначальним фактором. Для більш глибокого аналізу потрібно враховувати додаткові показники, такі як досвід, складність програми та результати на змаганнях.

### Практична частина

1. Виконайте дескриптивну статистику показників гравців з регбі у програмі MS Excel. Візуалізуйте отримані дані та зробіть висновок.

№ з/п	Вік	Довжина тіла, см	Маса тіла, кг	Кількість спіймань	Кількість приземлень
1.	24	188	95	28	14
2.	22	182	88	21	10
3.	25	190	102	32	18
4.	23	185	92	25	12
5.	26	192	98	30	16
6.	24	187	94	27	15
7.	23	184	90	24	13
8.	25	191	100	35	20
9.	22	180	86	19	8
10.	24	189	96	29	17

2. Виконайте дескриптивну статистику показників групи боксерів у програмі MS Excel. Візуалізуйте отримані дані та зробіть висновок.

№ з/п	Довжина тіла, см	Маса тіла, кг	Кількість перемог	Кількість нокаутів
1.	67,2	178	28	16
2.	70,5	182	22	12
3.	64,8	175	19	8
4.	72,1	185	31	20
5.	68,9	180	25	14
6.	66,3	177	18	10
7.	71,2	183	27	18
8.	69,7	181	23	15
9.	65,6	176	21	11
10.	73,4	187	29	19

3. Виміряні антропометричні дані дітей 9-10 років, що займаються легкою атлетикою. Виконайте дескриптивну статистику у програмі Statistica. Візуалізуйте отримані дані та зробіть висновок.

Довжина тіла, см	Маса тіла, кг	Обхват грудей, см	Довжина стопи, см
140	35	70	22
138	32	68	21
142	36	72	23
139	34	71	22
141	33	69	21
137	31	67	20
140	35	70	22
138	33	68	21
143	37	73	23
136	30	66	20

4. Виконайте дескриптивну статистику показників групи хокеїстів у програмі Statistica. Візуалізуйте отримані дані та зробіть висновок.

Довжина тіла, см	Маса тіла, кг	Очки (гольові передачі+голи)
185	92	48
178	85	32
190	98	62
182	88	41
188	94	55
183	90	39
179	86	35
192	100	67
177	83	28
186	93	51

## Практична робота № 5

### Тема «Параметричні методи перевірки статистичних гіпотез у програмах MS Excel та Statistica»

**Мета заняття:** ознайомитися із методами перевірки статистичних гіпотез та способами її розрахунку в програмах MS Excel та Statistica

#### Зміст і хід заняття

1. Використання інструменту «Пакет аналізу» в програмі MS Excel для виявлення відмінностей між вибірками.

2. Практична робота в системі Statistica.

Виконання та захист практичного завдання.

#### 1. Використання інструменту «Пакет аналізу» в програмі MS Excel для виявлення відмінностей між вибірками

**Приклад 1.** Виявити, чи достовірні відмінності при порівнянні даних частоти серцевих скорочень групи спортсменів до і після розминки. Таблиця з вхідними даними представлена на рисунку 25.

	А	В
1	до розминки	після розминки
2	70	135
3	66	126
4	61	115
5	72	140
6	80	121
7	59	112
8	64	130

Рис. 25. Таблиця з вхідними даними

#### Рішення

1. Введіть експериментальні дані до електронної таблиці.

2. Для виявлення достовірності відмінностей встановимо курсор в вільну комірку (наприклад, А11). Запустіть **Майстер функцій**, виберіть категорію **Статистичні** та **СТЮДЕНТ.ТЕСТ**. У діалоговому вікні функції **СТЮДЕНТ.ТЕСТ** введіть вхідні дані: в поле **массів1** введемо діапазон А2:А8; в поле **массів2** – діапазон даних досліджуваної групи В2:В8. В поле **Хвости** завжди вводиться з клавіатури цифра 2 (без лапок), а в полі **Тип** з клавіатури введемо цифру 1. Натиснемо кнопку **ОК**. В осередку А11 з'явиться значення ймовірності - 0,00000363.

3. Оскільки величина ймовірності випадкової появи аналізованих вибірок (0,00000363) менше рівня значущості ( $\alpha = 0,05$ ), то нульова гіпотеза відкидається. Отже, відмінності між вибірками не випадкові і середні вибірок вважаються такими, що достовірно відрізняються один від

одного. Тому на підставі застосування критерію Стьюдента можна зробити висновок про більшу ефективність виконаної розминки ( $p < 0,05$ ).

**Приклад 2.** При оцінці тренувального навантаження було визначено об'єм навантаження  $x_i$ , (год.), протягом місяця у 25 боксерів першої групи і 25 боксерів другої групи –  $y_i$ , (год.). На думку експертів, спортсмени з обох груп показали однаковий рівень досягнень на відбіркових змаганнях. Однак на здобуття цього обсягу вони витратили різне тренувальне навантаження. Оцінити цю відмінність на статистичну достовірність. Зробити висновок. Дані вимірювань представлені у таблиці 1.

Таблиця 1

Вхідні дані

№	$x_i$	$n_i$
1	70	4
2	75	5
3	82	10
4	86	1
5	88	3
6	90	2
Всього	-	25

№	$y_i$	$n_i$
1	90	3
2	95	3
3	98	7
4	100	4
5	105	3
6	110	5
Всього	-	25

Рішення

Для вирішення завдань такого типу використовуються так звані критерії відмінності, зокрема, **t-критерій Стьюдента**.

1. Введіть дані:  $x_i$  – в діапазон A2:A26;  $y_i$  – в діапазон B2:B26.

2. Вибір процедури здійснюється з трьох варіантів **t-тесту**. Необхідно розрахувати дисперсію для кожної групи. Встановіть курсор у вільну комірку (наприклад, A28). Увійдіть до меню **Майстер функцій**, виберіть категорію **Статистичні** та функцію **ДИСП.В**. У діалоговому вікні функції **ДИСП.В** введіть діапазон осередків A2:A26. Натисніть кнопку **ОК**. В осередку A28 з'явиться значення дисперсії – 41,3. Так само розрахуйте дисперсію для другої групи. Оскільки дисперсії двох груп нерівні. Для реалізації процедури в пункті меню **Дані** виберіть рядок **Аналіз даних** і далі вкажіть курсором миші на рядок **Двохвибірковий t-тест з різними дисперсіями**.

3. У діалоговому вікні задайте **Інтервал змінної 1**, вказуючи діапазон A1:A26.

4. Аналогічно вкажіть **Інтервал змінної 2**, тобто введіть посилання на діапазон другого стовпця B1:B26.

5. Поставте прапорець – *Мітки*.

6. Далі вкажіть вихідний діапазон. Для цього поставте перемикач в положення **Вихідний діапазон** і введіть в якості вихідного діапазону посилання на осередок C1. Клацніть по кнопці **ОК**. Результати аналізу. У

вихідному діапазоні C1:E13 з'являться результати процедури **Двохвибіркової t-тест з різними дисперсіями** (рис. 26).

Двохвыборочный t-тест с различными дисперсиями		
	$x_i$	$y_i$
Среднее	80,2	100,24
Дисперсия	41,33333333	40,69
Наблюдения	25	25
Гипотетическая разность средних	0	
df	48	
t-статистика	-11,06366493	
P(T<=t) одностороннее	4,17598E-15	
t критическое одностороннее	1,677224196	
P(T<=t) двухстороннее	8,35195E-15	
t критическое двухстороннее	2,010634758	

Рис. 26. Результат розрахунку Двохвибіркового t-тесту з різними дисперсіями

**Інтерпретація результатів.** Оскільки при порівнянні  $t$ -статистика  $>$   $t$  критичне двостороннє (11,1  $>$  2,02), різницю між вибірками слід вважати статистично достовірною. Це означає, що дві групи боксерів, показавши на змаганнях ідентичну кваліфікацію, витратили на її вдосконалення різний об'єм тренувального навантаження, а саме: перша група  $\bar{x} = 80$  (год.), друга –  $\bar{y} = 100$  (год.). Таким чином, тренувальне навантаження другої групи слід переглянути.

## 2. Практична робота в системі Statistica

**Приклад 3.** Порівняти різні за змістом тренувальні процеси контрольної групи  $x$  та експериментальної групи  $y$ . Нововведення міститься у процесі  $y$ . Показником якості тренувань служить час досягнення максимальної висоти підйому штанги під час підйому штанги на груди для поштовху, (с). У контрольній  $x$  та експериментальній  $y$  групах по 10 досліджуваних. Дані вимірювань представлені у таблиці 2.

Таблиця 2

### Вхідні дані

$x$	0,99	1,10	1,10	1,15	1,15	1,20	1,20	1,20	1,25	1,25
$y$	0,95	0,95	0,95	0,98	0,98	0,98	0,99	1,10	1,15	1,15

Рішення

1. Введіть експериментальні дані до електронної таблиці.
2. Відкрийте стартове вікно модуля **Basic Statistics/Tables – Основні статистики/Таблиці** (рис. 27).

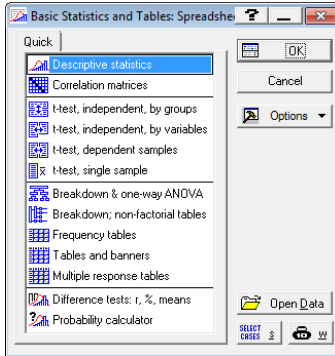


Рис. 27. Стартове вікно діалогу *t*-критерій у модулі Основні статистики/Таблиці системи STATISTICA

3. Клацніть у стартовому вікні модуля на рядку **t-test, independent, by variables** - *t*-критерій для незалежних вибірок. Наступне вікно з'явиться на моніторі (рис. 28).

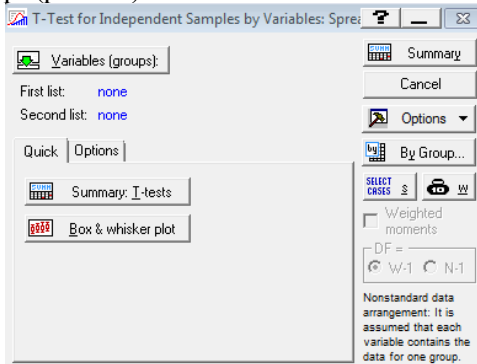


Рис. 28. Початкове вікно діалогу *t*-критерій для незалежних вибірок

4. Натисканням кнопки **Variables** – **Змінні** розкрийте вікно вибору змінних.

У лівому списку виберіть змінну (*x*), у правому – (*y*).

5. Далі клацніть **OK** або просто натисніть **Enter** на клавіатурі. У вікні знову натисніть кнопку **Summary**. На екрані з'явиться така таблиця (рис. 29):

T-test for Independent Samples (Spreadsheet2)											
Note: Variables were treated as independent samples											
	Mean	Mean	t-value	df	p	Valid N	Std.Dev.	Std.Dev.	F-ratio	p	
Group 1 vs. Group 2	Group 1	Group 2				Group 1	Group 2	Group 1	Group 2	Variances	Variances
x vs. y	1.159000	1.018000	3.896508	18	0.001058	10	10	0.079784	0.082030	1.057078	0.935457

Рис. 29. Таблиця результатів тестування (Приклад 3)



<b>Mean Group 1</b>	Середні арифметичні значення за групами
<b>Mean Group 2</b>	
<b>t-value</b>	Розрахункове значення критерію Стьюдента
<b>df</b>	Число ступенів свободи
<b>p</b>	Величина ймовірності випадкової появи аналізованих вибірок
<b>Std. Dev. Group 1</b>	Середнє квадратичне відхилення за групами
<b>Std. Dev. Group 2</b>	
<b>F-ratio Variances</b>	Розрахункове значення критерію Фішера
<b>p Variances</b>	Величина ймовірності випадкової появи аналізованих вибірок за критерієм Фішера

Найважливіше у цій таблиці рівень значущості  $p$  – той мінімальний рівень, на якому можна відкинути нульову гіпотезу.

У разі величина ймовірності випадкової появи аналізованих вибірок  $p=0,001058$  менше рівня значущості ( $\alpha = 0,05$ ).

Отже, оцінюючи зміст нового тренувального процесу за часом досягнення максимальної висоти підйому штанги при виконанні підйому штанги на груди для поштовху, приходимо до висновку про позитивний ефект нововведення.

**Приклад 4.** Оцінити у тенісистів зміни в показниках здатності до диференціації м'язових зусиль, (Н), в момент 1  $x_i$ ,  $y_i$  – в момент 2. Зробити висновок. Дані вимірювань представлені у таблиці 3.

Таблиця 3

Вхідні дані

$x_i$	2,3	2,3	2,4	2,4	2,4	2,5	2,5	2,5	2,6	2,7
$y_i$	2,7	3,0	3,0	3,5	3,5	3,5	3,9	3,9	3,9	4,2

Рішення

1. Введіть експериментальні дані до електронної таблиці.
  2. Відкрийте стартове вікно модуля **Basic Statistics/Tables – Основні статистики/Таблиці**.
  3. Клацніть у стартовому вікні модуля на рядку **t-test, dependent, samples - t-критерій для залежних вибірок**.
  4. Натисканням кнопки **Variables – Змінні** розкрийте вікно вибору змінних.
- У лівому списку виберіть змінну  $x_i$ , у правому –  $y_i$ .
5. Далі клацніть **OK** або натисніть **Enter** на клавіатурі. У вікні знову натисніть кнопку **Summary**. На екрані з'явиться така таблиця (рис. 30):

T-test for Dependent Samples (Spreadsheet16)								
Marked differences are significant at $p < ,05000$								
Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
$x_i$	2,460000	0,126491						
$y_i$	3,510000	0,484080	10	-1,05000	0,374907	-8,85656	9	0,000010

Рис. 30. Таблиця результатів тестування (Приклад 4)

Оскільки величина ймовірності випадкової появи аналізованих вибірок  $p = 0,000010$  менше рівня значущості ( $\alpha = 0,05$ ), можемо зробити висновок, що різниця між досліджуваними показниками у тенісистів статистично достовірна. Це означає, що у показниках диференціювання м'язових зусиль в обидві моменти часу спостерігаються істотні зміни.

### Практична частина.

1. У двох групах плавців  $x_i$  і  $y_i$  виміряна різниця частоти пульсу (уд./хв) між станом спокою і після максимально швидкого подолання дистанції. Встановити, чи достовірно відмінність за цими показниками в даних групах. Зробити висновок.

№	$x_i$	$n_i$
1	92	3
2	94	4
3	95	7
4	97	2
5	99	1
6	100	2
Всього	-	19

№	$y_i$	$n_i$
1	98	3
2	102	5
3	103	6
4	104	1
5	105	2
Всього	-	17

2. У таблиці представлені результати школярів з тесту «стрибок у довжину з місця» на початку  $x_i$  та наприкінці  $y_i$  навчального року. Встановити, чи достовірна відмінність за цими показниками. Визначити наскільки суттєво змінився досліджуваний показник протягом навчального року. Зробити висновок.

№	$x_i$	$y_i$	№	$x_i$	$y_i$
1	79	121	10	96	112
2	87	115	11	110	106
3	88	98	12	118	103
4	89	123	13	104	105
5	85	110	14	67	87
6	82	118	15	78	94
7	89	104	16	75	112
8	82	101	17	77	106
9	56	97			

## Практична робота № 6

### Тема «Непараметричні методи перевірки статистичних гіпотез в програмі Statistica»

**Мета заняття:** ознайомитися із методами перевірки статистичних гіпотез та способами її розрахунку в програмі Statistica

#### Зміст і хід заняття

1. Перевірка нормальності розподілу значень ознаки.
2. Практична робота із розрахунку непараметричних критеріїв перевірки гіпотез у програмі Statistica.

Виконання та захист практичного завдання.

#### 1. Перевірка нормальності розподілу значень ознаки.

Існує декілька способів перевірити, чи відповідають аналізовані дані закону нормального розподілу. Розглянемо два підходи, реалізовані у програмі Statistica.

Перший спосіб перевірки відповідності аналізованих даних нормальному розподілу.

На рисунку 31 представлені дані тестування школярів з тесту «стрибок у довжину з місця» на початку  $x_i$  та наприкінці  $y_i$  навчального року (Практична робота №5, Завдання 4). Необхідно встановити, чи розподілені ці дані за нормальним законом.

	1	2	3	4
	$X_i$	$Y_i$	Var3	Var.
1	79	121		
2	87	115		
3	88	98		
4	89	123		
5	85	110		
6	82	118		
7	89	104		
8	82	101		
9	56	97		
10	96	112		
11	110	106		
12	118	103		
13	104	105		
14	67	87		
15	78	94		
16	75	112		
17	77	106		
18				

Рис. 31. Результати школярів з тесту «стрибок у довжину з місця» на початку  $x_i$  та наприкінці  $y_i$  навчального року

У програмі Statistica є спеціальний модуль для перевірки відповідності даних тому чи іншому закону розподілу випадкових величин –

## Distribution Fitting – Підгонка розподілів. Цей модуль можна запустити з ГОЛОВНОГО МЕНЮ **Statistics** (рис. 32).

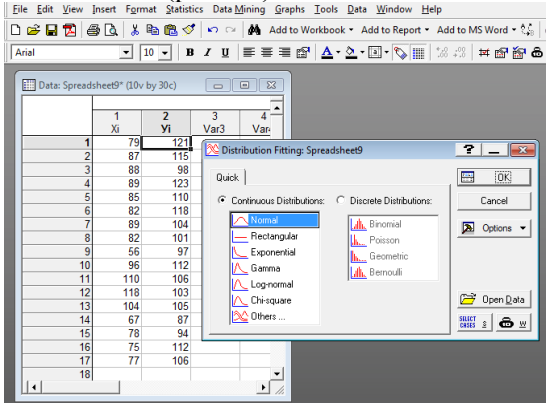


Рис. 32. Модуль *Distributions fitting* програми *Statistica*

Як видно на рисунку 32, у програмі *Statistica* можна порівнювати емпіричні розподіли з багатьма теоретичними законами розподілу випадкових величин. Необхідно перевірити, чи підпорядковуються дані про тестування школярів з тесту «стрибок у довжину з місця» на початку  $x_i$  та наприкінці  $y_i$  навчального року нормальному розподілу. Для цього у списку безперервних розподілів (**Continuous distributions**) вибираємо **Normal** і тиснемо **OK**. Далі з'явиться ще одне вікно (рис. 33), де необхідно вказати, яку саме змінну необхідно проаналізувати, і яким чином. Змінна для аналізу задається натисканням кнопки **Variables**. Інші налаштування можна залишити незмінними.

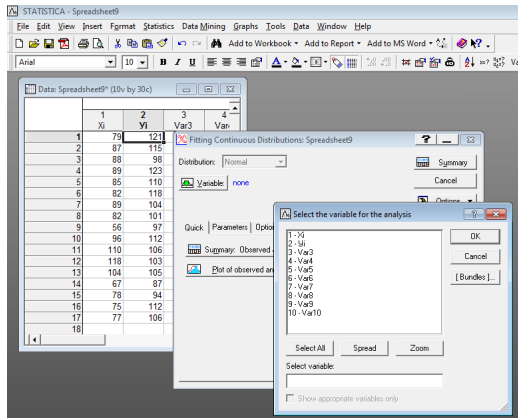


Рис. 33. Вікно *Fitting continuous distributions* (Підгонка безперервних розподілів) модуля *Distribution fitting*

Натиснувши на кнопку **Plot of observed and expected distributions – Зобразити спостережуваний та очікуваний розподіл**, отримаємо гістограму розподілу даних про результати тестування школярів з тесту «стрибок у довжину з місця» на початку  $x_i$  навчального року і дзвоноподібну червону криву (рисунок 34), відповідну очікуваному нормальному розподілу (у нього ті ж середні значення в аналізованій сукупності).

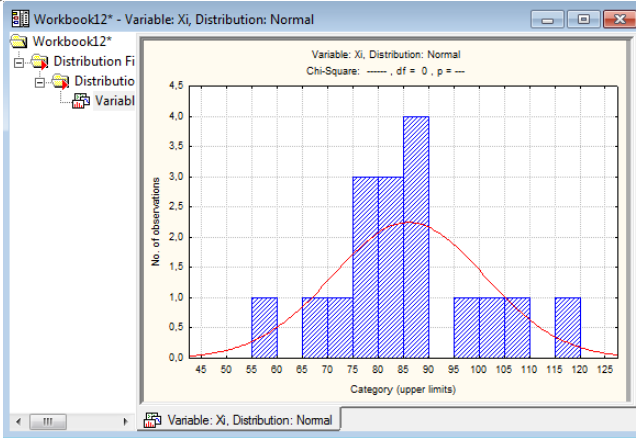


Рис. 34. Графічний результат аналізу, виконаного у модулі *Distribution Fitting*

В цілому розподіл значень аналізованої ознаки на рисунку збігається з нормальним (стовпчики гістограми приблизно вишиковуються в дзвоноподібну фігуру). Цей висновок заснований на візуальному аналізі розподілу.

Другий спосіб перевірки відповідності аналізованих даних нормальному розподілу.

В модулі **Descriptive Statistics – Описова статистика**, який знаходиться на вкладці: **Statistics > Basic Statistics/Tables**. Після запуску цього модуля необхідно відкрити закладку **Normality** і в полі **Distribution – Розподіл** обрати опції **Kolmogorov-Smirnov and Lilliefors test for normality – Тест Колмогорова-Смирнова та Лілієфорса на нормальність** та **Shapiro-Wilk's W test – W-тест Шаніро-Уїлка** (рисунок 35).

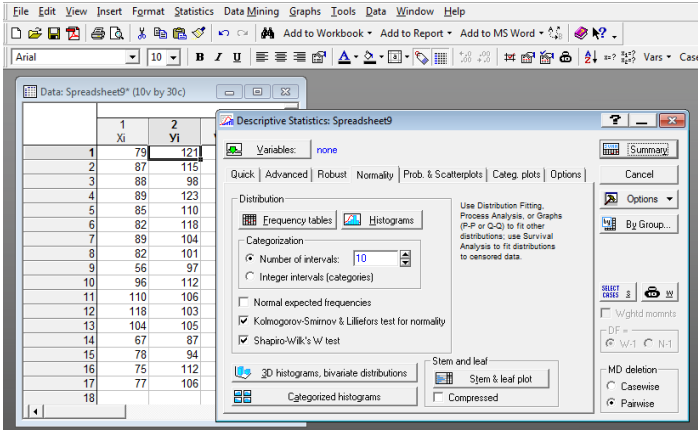


Рис. 35. Вікно модуля *Descriptive Statistics* на закладці *Normality*

Обидва ці тести перевіряють гіпотезу про відсутність відмінностей між емпіричним розподілом ознаки і теоретично очікуваним нормальним розподілом. Найбільш прийнятним є використання  $W$ -критерію Шапіро-Уилка, оскільки він має найбільшу потужність у порівнянні з усіма перерахованими критеріями (тобто частіше виявляє різницю між розподілами в тих випадках, коли вони дійсно є). Для вибору того чи іншого тесту треба поставити прапорець поруч з його назвою. Після вибору аналізованої змінної (кнопка **Variables**) та натискання кнопки **Histograms** буде побудовано гістограму розподілу значень ознаки та очікувану нормальну криву. Результати тестів на нормальність автоматично розміщуються у заголовку цього графіка. При  $p > 0,05$  слідує висновок про те, що аналізований розподіл не відрізняється від нормального.

## 2. Практична робота із розрахунку непараметричних критеріїв перевірки гіпотез у програмі Statistica.

### Приклад 1. (Критерій Вілкоксона)

У спринтера вимірювали швидкість бігу (м/с) початку  $x_i$  і в кінці  $y_i$  річного циклу тренувань. Оцінити характер змін швидкості бігу у спринтера за рік. Дані вимірювань представлені у таблиці 4.

## Вхідні дані

№	$x_i$	$y_i$	№	$x_i$	$y_i$
1	2,9	3	11	2,9	2,8
2	3	2,8	12	2,8	3
3	2,9	2,9	13	3,1	3,2
4	3,1	3,2	14	3,2	3,1
5	3,2	3,1	15	2,9	2,8
6	2,9	2,8	16	3	3,2
7	2,8	2,9	17	2,8	2,9
8	3	3	18	2,9	3
9	3,1	3,2	19	3,1	3,1
10	3,2	3,1	20	3,2	3

## Рішення

1. Введіть експериментальні дані до електронної таблиці.
2. Перевірте відповідність аналізованих даних нормальному розподілу. Якщо данні не відповідають закону нормального розподілу, необхідно використовувати непараметричні критерії.
3. Відкрийте вкладку **Statistics** стартове вікно модуля **Nonparametrics – Непараметричні**.
4. Клацніть у стартовому вікні модуля на рядку **Comparing two dependent samples variables – порівняння двох залежних змінних вибірки**.
5. Натисканням кнопки **Variables – Змінні** розкрийте вікно вибору змінних. У лівому списку виберіть змінну ( $x_i$ ), у правому – ( $y_i$ ).
6. Далі клацніть **OK** або натисніть **Enter** на клавіатурі.
7. Натисніть кнопку **Wilcoxon matched pairs test – Тест Вілкоксона** на відповідність пар. На екрані з'явиться така таблиця (рис. 36):

Wilcoxon Matched Pairs Test (Spreadsheet8)					
Marked tests are significant at p < ,05000					
Pair of Variables		Valid N	T	Z	p-value
Var4	& Var5	17	73,00000	0,165683	0,868406

Рис. 36. Таблиця результатів тестування (Приклад 1)

Оскільки величина ймовірності випадкової появи аналізованих вибірок  $p = 0,868406$  більше рівня значущості ( $\alpha = 0,05$ ), можемо зробити висновок, що різницю між порівнюваними вибірками слід вважати статистично недостовірною.

Оскільки відмінність швидкості бігу у випробуваного на початку  $x_i$  і наприкінці  $y_i$  річного циклу несуттєва. Це може бути пояснено

особливостями тренувального процесу у даному періоді (завданнями, спрямованістю навантажень тощо).

**Приклад 2.** Спортсмени двох окремих тренувальних груп пройшли тестування для визначення рівня гнучкості – визначали амплітуду нахилу (см). Порівняйте гнучкість спортсменів першої і другої груп. Дані вимірювань представлені у таблиці 5.

Таблиця 5

Вхідні дані										
$x_i$	20	31	31	32	32	32	35	36	38	
$y_i$	20	31	31	32	32	32	35	36	38	38

Рішення

1. Введіть експериментальні дані до електронної таблиці.
2. Перевірте відповідність даних закону нормального розподілу. Якщо данні не відповідають закону нормального розподілу, необхідно використовувати непараметричні критерії.
3. Відкрийте вкладку **Statistics** стартове вікно модуля **Nonparametrics – Непараметричні**.
4. Клацніть у стартовому вікні модуля на рядку **Comparing two independent samples (group) – порівняння двох незалежних вибірок**.
5. Натисканням кнопки **Variables – Змінні** розкрийте вікно вибору змінних. У лівому списку виберіть змінну ( $x_i$ ), у правому – ( $y_i$ ).
6. Далі клацніть **OK** або натисніть **Enter** на клавіатурі.
7. Натисніть кнопку **Mann-Whitney U test – Критерій Манна-Уїтні**.

На екрані з'явиться така таблиця (рис. 37):

Mann-Whitney U Test (Spreadsheet9)									
By variable $y_i$									
Marked tests are significant at $p < .05000$									
variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2
$x_i$			0,00	0,00	1,000000	0,00	1,000000	1	2

Рис. 37. Таблиця результатів тестування (Приклад 2)

Оскільки величина ймовірності випадкової появи аналізованих вибірок  $p = 1,000000$  більше рівня значущості ( $\alpha = 0,05$ ), можемо зробити висновок, що відмінність між вибірками є статистично недостовірною. Встановлено, що спортсмени обох груп не відрізняються за рівнем розвитку гнучкості хребта.



### Практична частина.

1. Порівняти показники витривалості групи велосипедистів до тренувань  $x_i$  і після  $y_i$ , (км). Визначити чи є відмінності між даними показниками. Зробити висновок. (за критерієм Вілкоксона).

№	$x_i$	$y_i$	№	$x_i$	$y_i$
1	327,4	330,5	7	395	395
2	330,5	352,4	8	398,5	402
3	350,2	360,7	9	399,5	385
4	338,1	344,4	10	399,9	402
5	360,9	358,2	11	405	408,2
6	381,4	390,5	12	410	411,5

2. У 32 спортсменів виміряна відносна сила розгиначів ніг (сила обох ніг в сумі в ньютонах на 10 кг маси тіла спортсмена) на початку підготовчого періоду  $x_i$  і в кінці  $y_i$ . Встановити, чи достатньо ефективно змінилася відносна сила розгиначів ніг у цих спортсменів протягом підготовчого періоду. Зробити висновок. (за критерієм знаків)

№	$x_i$	$y_i$	№	$x_i$	$y_i$
1	7,20	7,20	17	9,10	9,05
2	7,30	7,35	18	9,40	4,45
3	7,40	7,45	19	9,50	9,60
4	7,50	7,60	20	9,50	9,65
5	7,90	7,80	21	9,60	9,70
6	8,10	8,00	22	9,60	9,60
7	8,20	8,30	23	9,60	9,60
8	8,30	8,35	24	9,60	9,80
9	8,40	8,50	25	9,70	9,65
10	8,50	8,50	26	9,70	9,70
11	8,60	8,65	27	9,80	9,85
12	8,90	8,90	28	9,80	9,80
13	8,90	8,95	29	9,80	9,75
14	9,00	9,10	30	9,80	9,80
15	9,10	9,15	31	9,90	10,00
16	9,10	9,10	32	9,90	9,90

3. У рівноцінних умовах у двох плавців  $x_i$  і  $y_i$  10 разів фіксувався середній час (с) пропливання 25-метрової дистанції. Порівняйте можливості плавців. (за критерієм Манна-Уїтні)

$x_i$	12,4	12,5	12,3	12,8	12,5	11,0	12,2	12,4	13,0	12,7
$y_i$	12,8	11,0	12,5	12,4	12,7	13,0	12,8	13	12,5	12,2

## Практична робота № 7

### Тема «Кореляційний аналіз із використанням параметричних методів у програмах MS Excel та Statistica»

**Мета заняття:** ознайомитися із методами визначення взаємозв'язку між ознаками та способами розрахунку в програмах MS Excel та Statistica

#### Зміст і хід заняття

1. Практична робота із розрахунку коефіцієнту кореляції в програмі MS Excel.

2. Практична робота із розрахунку коефіцієнту кореляції в програмі Statistica.

Виконання та захист практичного завдання.

#### 1. Практична робота із розрахунку коефіцієнту кореляції в програмі MS Excel.

**Приклад 1.** Встановити залежність між швидкістю вильоту шайби  $x_i$ , (м/с), і швидкістю ковзання гравців на льоду  $y_i$ , (м/с). Зробити висновок. Таблиця з вхідними даними представлена на рисунку 38.

	A	B
1	$x_i$	$y_i$
2	20,0	4,0
3	21,2	4,3
4	23,4	4,4
5	24,0	4,2
6	25,0	4,3
7	25,5	4,4
8	26,0	4,3
9	26,1	4,5
10	27,0	4,4
11	27,3	4,6


Рис. 38. Таблиця з вхідними даними

#### Рішення

Для виявлення ступеня взаємозв'язку насамперед необхідно ввести дані до електронної таблиці. Потім обчислюється значення коефіцієнта кореляції між вибірками. Для цього встановіть курсор в вільну комірку (наприклад, A13). Увійдіть до меню **Майстер функцій**, оберіть функцію **КОРРЕЛ**. Введіть в поле **масив1** діапазон даних A2:A11. В поле **масив2** введіть діапазон даних B2:B11. Натисніть кнопку **ОК**. В осередку A13 з'явиться значення коефіцієнта кореляції – 0,775071. Значення коефіцієнта кореляції знаходиться в межах від 0,70 до 0,90. Отже, в дослідженій групі хокеїстів спостерігається тісна кореляційна залежність між швидкістю вильоту шайби та швидкістю їх ковзання на льоду. Таким чином, робимо

висновок, що зі збільшенням швидкості ковзання на льоду швидкість вильоту шайби буде більшою.

За даними вибірок побудувати графік кореляційного поля.

1. Виділіть діапазон осередків A2:B11.
2. Виберіть команду **Вставка**> **Діаграми**> **Точкова** .
3. Помістіть діаграму на вільне місце на аркуші.
4. Натисніть правою кнопкою миші на *горизонтальній осі (значень)*.
5. Із контекстного меню виберіть **Формат осі**.
6. У вікні **Формат осі** змініть мінімальне значення на 19.
7. Використовуючи **Елементи діаграми** додайте **Лінію тренду**.
8. Для даного прикладу отримаємо наступний графік (рис. 39):

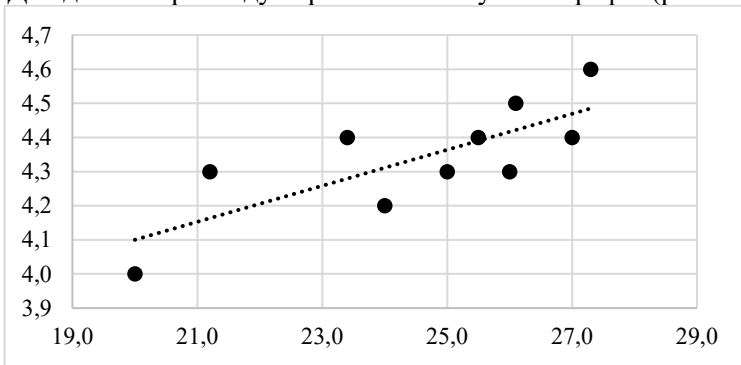


Рис. 39. Графік кореляційного поля

9. Аналіз графіку дозволяє зробити висновок, що в дослідженій групі хокеїстів спостерігається прямий лінійний взаємозв'язок між швидкістю вильоту шайби та швидкістю їх ковзання на льоду. Оскільки точки групуються всередині еліпса правильної форми (ось його є прямою). Це лінійна залежність. Силу зв'язку можна оцінити за допомогою коефіцієнта кореляції.

Особливістю графічного способу оцінки кореляційного зв'язку є наочність.

Загалом за графіком можна визначити такі моменти:

- якщо експериментальні точки розсіяні по полю графіка хаотично і по них неможливо провести лінію, то кореляція відсутня;
- якщо точки групуються вздовж будь-якої лінії, то кореляція є, і вона тим тісніша, чим щільніше розташовані ці точки.
- по напрямку лінії, вздовж якої групуються точки, можна визначити вид кореляції (позитивна чи негативна).

**Приклад 2.** Проаналізувати результати тестування спеціальної підготовленості борців – визначити наявність і характер взаємозв'язку між показниками. Таблиця з вхідними даними представлена на рисунку 40.

	A	B	C	D
	Час човникового бігу 4*9 (с)	Час виконання 18 кидків (с)	Підтягування на перекладині 10 раз (с)	Підтягування на перекладині (кількість разів)
1				
2	10,5	41	10	27
3	10,4	38	9	25
4	9,9	34	9	30
5	10,1	36	8	27
6	9,9	30	8	32
7	10	34	8	32
8	10	35	10	24
9	10	38	8	28
10	10,2	34	9	27
11	10,1	36	9	24

Рис. 40. Таблиця з вхідними даними

### Рішення

Виберіть команду **Дані – Аналіз даних** і виберіть рядок **Кореляція**. У діалоговому вікні вкажіть **Вхідний інтервал** A1:D11. Вкажіть, що дані розглядаються по рядках. Вкажіть вихідний діапазон. Для цього поставте прапорець в ліве поле **Вихідний інтервал** і в праве поле введення **Вихідний інтервал** введіть A14. Натисніть кнопку **ОК**. Результати аналізу.

*Результати аналізу.* У вихідному діапазоні отримуємо кореляційну матрицю (рис. 41).

	Час човникового бігу 4*9 (с)	Час виконання 18 кидків (с)	Підтягування на перекладині 10 раз (с)	Підтягування на перекладині (кількість разів)
Час човникового бігу 4*9 (с)	1			
Час виконання 18 кидків (с)	0,7600797	1		
Підтягування на перекладині 10 раз (с)	0,500870323	0,433576554	1	
Підтягування на перекладині (кількість разів)	-0,475958131	-0,536566486	-0,610882839	1

Рис. 41. Результати обчислення кореляційної матриці з прикладу 2

*Інтерпретація результатів.* З таблиці видно, що найбільші коефіцієнти кореляції спостерігаються між часом виконання човникового бігу і 18 кидків ( $r = 0,76$ ), підтягуванням на перекладині ( $r = 0,50$ ), часом виконання 18 кидків і 10 підтягувань на перекладині ( $r = 0,43$ ). Негативна кореляція середньої сили виявлена між кількістю підтягувань на перекладині і часом виконання 10 підтягувань ( $r = -0,61$ ), часом виконання 18 кидків ( $r = -0,54$ ), часом виконання човникового бігу і кількістю підтягувань на перекладині ( $r = -0,48$ ). Мається на увазі, що в порожніх клітинах в правій верхній половині таблиці знаходяться ті ж коефіцієнти кореляції, що і в нижній лівій (симетрично розташовані щодо діагоналі).

#### 4. Практична робота в системі Statistica

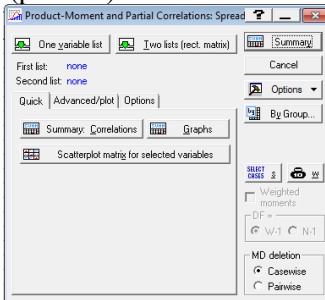
**Приклад 3.** Визначити, чи існує залежність між зростанням важкоатлетів  $x_i$ , (см), і максимальною швидкістю руху штанги у фазі тяги при виконанні ривка  $y_i$ , (м/с), при інших рівних умовах. Зробити висновок. Дані вимірювань представлені у таблиці 6.

Таблиця 6

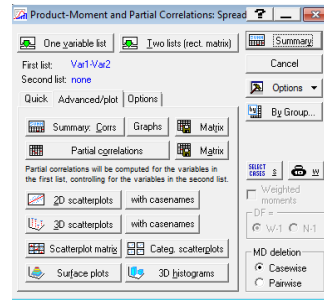
Вхідні дані										
$x_i$	165	166	169	173	177	178	179	179	180	182
$y_i$	1,3	1,3	1,3	1,3	1,3	1,4	1,4	1,4	1,4	1,4

Рішення

1. Введіть експериментальні дані до електронної таблиці.
2. Відкрийте стартове вікно модуля **Basic Statistics/Tables – Основні статистики/Таблиці**.
3. Клацніть у стартовому вікні модуля на рядку **Correlation matrices – Кореляційні матриці**. Наступне вікно з'явиться на моніторі (рис. 42).



а)



б)

Рис. 42. Початкове вікно діалогу Кореляційні матриці

Як і більшість діалогових вікон аналізу, вікно **Кореляційні матриці** містить набір опцій. Як правило, пропонується не менш як два типи аналізу.

Вкладка **Quick – Швидко** цього вікна містить опції, що дозволяють обчислити парні кореляції декількох змінних і відобразити результати в різних формах, наприклад вивести рівень значущості тощо (рис. 42 а)

Вкладка **Advanced/plot – Розширений/сюжет** містить опції для обчислення як парних, так і приватних кореляцій, а також опції збереження матриць, отримання деяких статистик або графіків. Залежно від вибраного аналізу можуть бути доступні також додаткові вкладки (рис. 42 б).

За допомогою двох кнопок угорі вікна задаються змінні для двох можливих типів кореляційних матриць: **One variable list** – (квадратної матриці) та **Two list (rest. matrix)** – (прямокутної матриці).

4. Клацніть правою кнопкою **Two list (rest. matrix)**– **Прямокутна матриця (два списки)** для відкриття вікна вибору змінних. Виберіть змінні та натисніть **ОК**. У вікні знову натисніть кнопку **Summary**. Система здійснить обчислення, і побачите кореляційну матрицю на екрані (рис. 43).

Correlations (Spreadsheet11)	
Marked correlations are significant at p < ,05000	
N=10 (Casewise deletion of missing data)	
Variable	yi
xi	0.870521

Рис. 43. Таблиця кореляцій для змінних із двох обраних списків

У цій матриці є лише один стовпець, оскільки у другому списку обрали лише одну змінну. У стовпці дано коефіцієнт кореляції між змінною  $x_i$  та змінною  $y_i$ .

Знайдений коефіцієнт кореляції  $r_{xy} = 0,87$  вказує на тісний позитивний кореляційний взаємозв'язок між зростанням важкоатлетів та максимальною швидкістю руху штанги у фазі тяги при виконанні ривка.

5. Поверніться до початкового вікна діалогу **Кореляційні матриці**, виберіть вкладку **Advanced/plot** – **Розширений/сюжет**, натисніть кнопку **2D scatterplot / 2М діаграма розсіювання** та **ОК**. Наступний графік з'явиться на екрані (рис. 44).

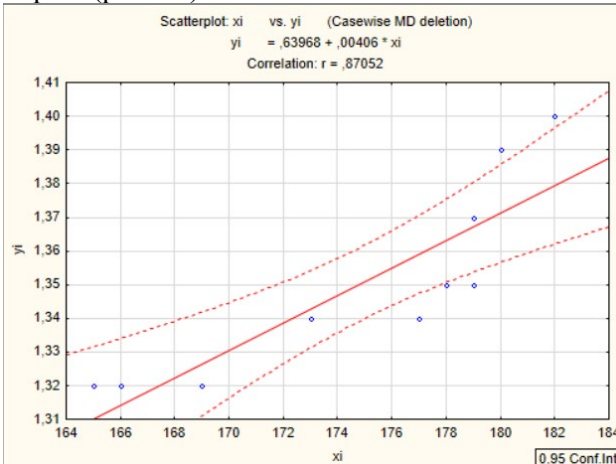


Рис. 44. Діаграма розсіювання з підігнаною лінійною регресією та довірчою смугою

*Налаштування графіка.* Зауважте, що при побудові активного графічного вікна з'явилася відповідна панель інструментів. Панель **Графічних інструментів** відрізняється від панелей інструментів електронних таблиць (рис. 45).

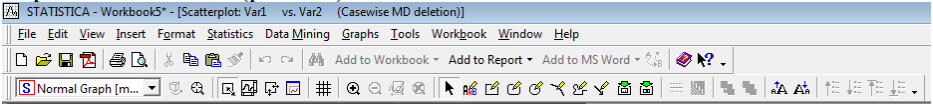


Рис. 45. Панель Графічних інструментів

Вона містить значну кількість можливостей налаштування графіка та інструментів для малювання. Всі ці можливості також доступні у меню, і багато з них – з контекстних меню, що з'являються при натисканні правою кнопкою миші на відповідній частині графіка.

Зауважимо, що опції контекстного меню розміщуються ієрархічно. Перші одна або дві опції відносяться лише до виділеного елемента графіку, тоді як нижні опції відносяться до елементів, які так чи інакше пов'язані з ним.

**Приклад 4.** У таблиці наведено антропометричні дані групи дітей із групи початкової підготовки. Визначити наявність і характер взаємозв'язку між досліджуваними ознаками. Обчислити кореляційну матрицю. Зробити висновок. Дані вимірювань представлені у таблиці 7.

Таблиця 7

Вхідні дані

Обхват голови, см	Обхват грудей, см	Обхват плечей, см	Обхват таза, см	Довжина тіла, см	Довжина тіла (сидячі), см	Довжина рук, см	Довжина ніг, см	Маса тіла, кг
55	67	30	23	125	110	53	62	28
53	67	28	22	134	113	57	69	28
54	66	31	20	123	108	53	62	26
54	69	31	22	142	116	59	71	33
54	67	31	22	133	114	55	72	31
55	72	33	22	133	111	57	69	29
50	63	24	19	124	110	54	64	24
53	62	29	21	130	110	57	66	25
53	62	30	22	132	112	57	68	32
55	66	29	23	124	108	53	62	27
53	67	28	22	124	110	54	64	24
54	64	31	20	130	110	56	66	25
53	69	31	22	132	112	57	68	32
54	67	31	22	134	113	57	69	28
50	63	24	19	133	114	55	72	31

Рішення

1. Введіть експериментальні дані до електронної таблиці.

2. Відкрийте стартове вікно модуля **Basic Statistics/Tables – Основні статистики/Таблиці**.

3. Клацніть у стартовому вікні модуля на рядку **Correlation matrices – Кореляційні матриці**.

Натисніть кнопку **One variable list – Квадратна матриця (один список)**. У вікні вибору змінних виберіть усі змінні, тобто. натисніть кнопку **Select All – Вибрати все**, потім натисніть **OK**. У вікні знову натисніть кнопку **Summary**. Наступна таблиця з'явиться на екрані (рис. 46).

Correlations (Spreadsheet3)											
Marked correlations are significant at $p < .05000$											
N=15 (Casewise deletion of missing data)											
Variable	Means	Std Dev.	Обхват голови, см	Обхват грудей, см	Обхват плечей, см	Обхват таза, см	Довжина тіла, см	Довжина тіла (сидячі), см	Довжина рук, см	Довжина ніг, см	Маса тіла, кг
Обхват голови, см	53.3333	1.543033	1.000000	0.565017	0.868669	0.748728	0.034975	-0.201893	0.024593	-0.218259	0.075839
Обхват грудей, см	66.0667	2.840188	0.565017	1.000000	0.595818	0.534616	0.288820	0.204015	0.205764	0.171226	0.303248
Обхват плечей, см	29.4000	2.557901	0.868669	0.595818	1.000000	0.572108	0.278498	0.007307	0.332322	0.058459	0.272665
Обхват таза, см	21.4000	1.298351	0.748728	0.534616	0.572108	1.000000	0.143403	0.062385	0.128604	-0.056029	0.302841
Довжина тіла, см	130.2000	5.294202	0.034975	0.288820	0.278498	0.143403	1.000000	0.899125	0.890259	0.879627	0.726773
Довжина тіла (сидячі), см	111.4000	2.292846	-0.201893	0.204015	0.007307	0.062385	0.899125	1.000000	0.685205	0.911266	0.743111
Довжина рук, см	55.6000	1.882248	0.024893	0.205764	0.332322	0.128604	0.890259	0.685205	1.000000	0.704250	0.512292
Довжина ніг, см	66.9333	3.534860	-0.218259	0.171226	0.058459	-0.056029	0.879627	0.911266	0.704250	1.000000	0.696531
Маса тіла, кг	28.2000	3.051931	0.075839	0.303248	0.272665	0.302841	0.726773	0.743111	0.512292	0.696531	1.000000

Рис. 46. Таблиця з кореляційною матрицею для вибраних змінних

4. Розглянемо докладніше цю матрицю. У ній представлені коефіцієнти кореляції між досліджуваними ознаками. Виберіть у стовпці та рядку імена змінних, на перетині рядка та стовпця знаходиться потрібний коефіцієнт кореляції, наприклад, виберіть Обхват голови та Обхват таза – на перетині стоїть число 0,748728 – коефіцієнт кореляції між цими змінними.

Червоним кольором автоматично виділені коефіцієнти, значущі на рівні  $p < 0,05$ . Саме на ці коефіцієнти слід звернути особливу увагу.

5. Щоб побудувати діаграму розсіювання для цих змінних, клацніть правою кнопкою миші відповідний коефіцієнт кореляції (0,748728). У контекстному меню виберіть пункт **Графіки вихідних даних**, а далі потрібний тип графіка, як показано нижче (рис. 47).

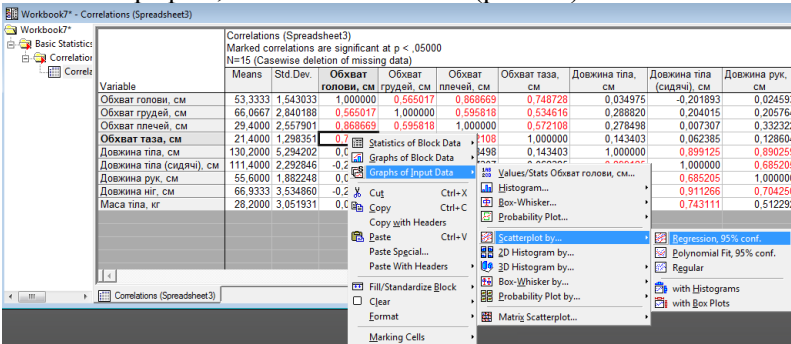


Рис. 47. Алгоритм побудови графіка вихідних даних



Потрібний графік буде відображено на екрані (рис. 48).

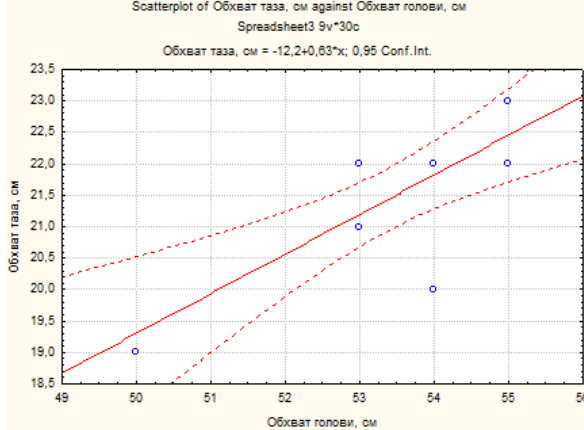


Рис.48. Графічне подання залежності параметрів, що вивчаються

### Практична частина.

1. Встановити взаємозв'язок між кількістю кидків за 10 хв. з різними противниками  $x_i$  і частотою серцевих скорочень  $y_i$ , (уд./хв.) у 10 спортсменів. Зробити висновок.

$x_i$	5	6	7	7	8	8	8	9	9	9
$y_i$	155	160	160	160	165	170	170	175	185	185

2. У таблиці наведено антропометричні дані і результати виконання тестів на гнучкість хлопчиків 6-ти років. Визначити існуючі взаємозв'язки між досліджуваними ознаками. Обчислити кореляційну матрицю. Зробити висновок.

Обхват, ширина				Довжина тіла		Маса тіла	Гнучкість		
голови, см	грудей, см	плечей, см	таза, см	рук, см	ніг, см	маса тіла, кг	нахил тулуба вперед з положення сидячи, см	місток з положення лежачи, см	поперечний або повздовжній шпатель, см
1	2	3	4	5	6	7	8	9	10
51	57	29	22	53	62	23	13	39	3
52	61	28	18	51	65	22	-2	57	18
52	62	27	20	49	59	22	16	35	7

1	2	3	4	5	6	7	8	9	10
53	62	27	20	48	65	23	4	51	12
54	58	25	18	48	64	21	-7	48	13
54	58	25	18	48	64	22	-5	52	15
53	62	27	20	48	65	23	4	51	12
52	61	28	18	51	65	22	-2	57	18
52	62	27	20	49	59	22	16	35	7
52	59	26	19	48	63	22	-7	48	13
54	65	26	20	48	65	23	4	51	10
54	65	26	20	50	64	25	-13	61	15
52	62	26	19	49	59	22	-5	52	15
51	57	29	22	50	64	25	1	51	13
54	58	25	18	48	63	21	4	50	12
52	61	28	19	51	65	22	0	54	18
52	62	26	19	49	59	22	17	35	9
52	57	29	22	50	64	25	13	37	3

## Практична робота № 8

### Тема «Використання непараметричних методів в кореляційному аналізі»

**Мета заняття:** ознайомитися із непараметричними методами визначення взаємозв'язку між ознаками та способу розрахунку в програмі Statistica

#### Зміст і хід заняття

1. Коефіцієнт кореляції (Спірмена R, тау Кендала, Гамма).

Виконання та захист практичного завдання.

#### 1. Коефіцієнт кореляції (Спірмена R, тау Кендала, Гамма)

**Приклад 1.** 10 акробатів, які посіли на змаганнях місця  $x_i$ , виконують спеціальний тест на витривалість. За результатами тесту на витривалість спортсмени розподіляються по рангу  $y_i$ . Визначити, чи впливає спеціальна витривалість на змагальний результат. Дані вимірювань представлені у таблиці 8.


Таблиця 8

#### Вхідні дані

$x_i$	1	2	3	4	5	6	7	8	9	10
$y_i$	3	2	1	5	4	6	8	7	10	9

Рішення

1. Введіть експериментальні дані до електронної таблиці.
2. Відкрийте стартове вікно модуля **Statistics** → **Nonparametrics** або

натисніть кнопку  на додатковій панелі інструментів → **Correlations (Spearman, Kendall tau, gamma)** (Кореляції (Спірмена, тау Кендала, Гама) → ОК (рис. 49).

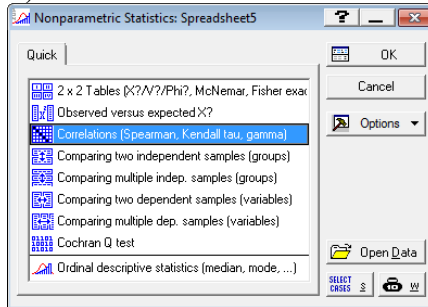


Рис. 49. Діалогове вікно Непараметричні статистики та розподіли

3. У діалоговому вікні модуля **Nonparametric**, що відкрилося. **Correlations (Непараметричні кореляції)** натисніть кнопку **Variables** і виберіть змінні  $x_i$  та  $y_i$ . Натисніть кнопку **OK**. Натисніть кнопку **Spearman R**. На екрані з'явиться така таблиця (рис. 50).

Variable	$x_i$	$y_i$			
$x_i$	1,000000	0,915152			
$y_i$	0,915152	1,000000			

Рис. 50. Вікно з результатами аналізу даних

Коефіцієнт кореляції  $r_{xy} = 0,915152$  вказує на те, що у досліджених акробатів досить тісний зв'язок між результатами їх виступів на змаганнях і виконаним тестом на витривалість.

Крім коефіцієнта кореляції Спірмена, модуль **Nonparametric Correlations** дозволяє проводити розрахунки коефіцієнтів Кендала і Гама. Коефіцієнт Кендалл рекомендується використовувати, якщо хоча б одна змінна є якісною (порядковою), коефіцієнт гама – якщо змінні містять багато значень, що повторюються.

4. Статистики **Кендала тау** і **Гама**. Поверніться у вікно **Непараметричні кореляції** та виберіть опцію Статистика Тау Кендала, а потім опцію Гама (рис. 51, 52).

		Kendall Tau Correlations (Spreadsheet82)			
		MD pairwise deleted			
		Marked correlations are significant at $p < .05000$			
Variable		Var1	Var2		
Var1		1,000000	0,733333		
Var2		0,733333	1,000000		

Рис. 51. Результати розрахунків статистики Кендалла тау

		Gamma Correlations (Spreadsheet82)			
		MD pairwise deleted			
		Marked correlations are significant at $p < .05000$			
Variable		Var1	Var2		
Var1		1,000000	0,733333		
Var2		0,733333	1,000000		

Рис. 52. Результати розрахунків статистики Гамма

Ці критерії тісно пов'язані між собою, але відрізняються від статистики Спірмена. Коефіцієнт кореляції Спірмена  $R$  можна представити як обчислену по рангам кореляцію Пірсона. Статистики Кендала тау і Гамма швидше оцінюють ймовірності. Різницю між ймовірністю того, що значення змінних, що спостерігаються, мають один і той же порядок, і ймовірністю того, що порядок різний.

### Практична частина.

1. Група експертів оцінила техніку захисту у 10 боксерів  $x_i$ , бали (під прямий удар правою в голову підставляється ліве плече). В ході боїв експерти оцінили дії цих боксерів в захисті  $y_i$ , бали. Оцінювали за наступною шкалою – 5 відповідає найкращому результату, 2 – найгіршому. Встановити, як техніка захисту впливає на ефективність її застосування у боксерському поєдинку.

№	$x_i$	$y_i$	№	$x_i$	$y_i$
1	5	4	6	2	3
2	4	5	7	5	4
3	5	5	8	4	4
4	4	4	9	3	2
5	3	4	10	2	3

2. Дві групи важкоатлетів одного віку, статі і кваліфікації займаються за різними програмами розвитку сили. В обох групах перед змаганнями виміряна станова сила спортсменів, відповідно до чого вони розподілилися по місцях  $x_i$ . В ході змагань спортсмени розподілили місця  $y_i$ . Встановити, чи існує зв'язок між рівнем розвитку станової сили і результатом у змаганні.

1 група

$x_i$	$y_i$
1	3
2	2
3	1
4	6
5	5
6	4
7	8
8	7
9	10
10	9

2 група

$x_i$	$y_i$
1	4
2	3
3	2
4	1
5	6
6	8
7	7
8	5
9	10
10	9

3. Встановити наявність і характер взаємозв'язку між результатом оцінки інтелекту за тестом Векслера та оцінкою з математики. Зробити висновок.

№	Бал тесту Векслера	Оцінка з матем.	№	Бал тесту Векслера	Оцінка з матем.
1	5	4	10	15	7
2	7	5	11	16	8
3	10	9	12	10	8
4	10	10	13	10	6
5	15	11	14	8	7
6	12	8	15	6	5
7	16	10	16	11	8
8	15	8	17	10	9
9	18	11			

### Практична робота № 9

#### Тема «Дисперсійний аналіз у програмах MS Excel та Statistica»

**Мета заняття:** ознайомитися із методами перевірки статистичних гіпотез та способами розрахунку в програмах MS Excel та Statistica

#### Зміст і хід заняття

1. Практична робота із виконання однофакторного дисперсійного аналізу в програмі MS Excel.
2. Практична робота із виконання однофакторного дисперсійного аналізу в програмі Statistica.

Виконання та захист практичного завдання.

#### 1. Практична робота із виконання однофакторного дисперсійного аналізу в програмі MS Excel.

**Приклад 1.** Виміряно час читання (у секундах) одного і того ж тексту у респондентів різних вікових груп: до 22 років, від 23 до 26 років, від 27 років. Результати виміру представлені у таблиці 9.

Вхідні дані

№ респондента	Вік	Час	№ респондента	Вік	Час	№ респондента	Вік	Час
1	20	35	1	23	41	1	27	46
2	20	37	2	23	40	2	27	40
3	21	43	3	24	42	3	28	40
4	21	35	4	24	39	4	28	39
5	21	35	5	24	38	5	28	38
6	21	39	6	25	37	6	29	37
7	21	40	7	25	37	7	29	60
8	22	34	8	25	38	8	29	39
9	22	34	9	26	40	9	30	40
10	22	34	10	26	41	10	31	38
11	22	35	11	26	39			
12	22	36						
13	22	37						

Потрібно встановити рівень статистичної значущості відмінностей середніх значень часу читання текстів респондентами різних вікових груп: до 22 років, від 23 до 26 років, від 27 років.

Фактором є вік респондентів.

Фактор має три градації: до 22 років, від 23 до 26 років, від 27 років.

Досліджуваною змінною (властивістю) є час, витрачений респондентом на читання тексту. Замінімо в таблиці стовпець «Вік» на стовпець «Фактор» (табл. 10).

Таблиця 10

Вхідні дані

№	Фактор	Час	№	Фактор	Час	№	Фактор	Час
1	20	35	1	23	41	1	27	46
2	20	37	2	23	40	2	27	40
3	21	43	3	24	42	3	28	40
4	21	35	4	24	39	4	28	39
5	21	35	5	24	38	5	28	38
6	21	39	6	25	37	6	29	37
7	21	40	7	25	37	7	29	60
8	22	34	8	25	38	8	29	39
9	22	34	9	26	40	9	30	40
10	22	34	10	26	41	10	31	38
11	22	35	11	26	39			
12	22	36						
13	22	37						

1. На окремому аркуші в Excel створіть таблицю зі стовпцями: до 22 років, від 23 до 26 років, від 27 років (назви градацій фактору). З протоколу до створеної таблиці перенесіть відповідні значення змінної (час, витрачений на читання тексту).

2. Знайдіть у групі команд Дані команду **Аналіз даних**. Натисніть кнопку **ОК**.

3. В **Інструментах аналізу** знайдіть **Однофакторний дисперсійний аналіз**. Натисніть кнопку **ОК**.

4. У вікно **Вхідний інтервал** внесіть коди створеної таблиці, виділивши цю таблицю.

5. У вікні **Мітки у першому рядку** встановіть прапорець.

6. В області **Параметри виводу** слід вибрати режим **Вихідний інтервал** – після цього в доступному вікні ввести (або вказати клацанням миші на аркуші) адресу комірки E1. Натисніть кнопку **ОК**.

7. Починаючи з осередка E1, будуть виведені результати обробки – для наочності представлені на рис. 53.

	E	F	G	H	I	J	K
Однофакторный дисперсионный анализ							
ИТОГИ							
	<i>Группы</i>	<i>Счет</i>	<i>Сумма</i>	<i>Среднее</i>	<i>Дисперсия</i>		
	до 22 років	13	474	36,46154	7,435897436		
	від 23 до 26 років	11	432	39,27273	2,818181818		
	від 27 років	10	417	41,7	47,34444444		
Дисперсионный анализ							
	<i>Источник вариации</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-Значение</i>	<i>F критическое</i>
	Между группами	157,2227	2	78,61135	4,4837084	<b>0,02</b>	3,304817252
	Внутри групп	543,5126	31	17,53266			
	Итого	700,7353	33				

Рис. 53. Результат роботи інструменту Однофакторний дисперсійний аналіз

У стовпці *P-Значення* зазначено рівень значущості відмінностей:  $\alpha = 0,02$  (число округлено до двох знаків після коми).

**Статистичний висновок.** Оскільки  $\alpha = 0,02$ , тобто,  $p < 0,05$ , приймається гіпотеза  $H_1$ : Середні значення принаймні двох вибірок статистично значуще розрізняються ( $p < 0,05$ ).

**Педагогічний висновок.** Виявлено статистично значущі ( $p < 0,05$ ) відмінності середніх значень часу, витраченого на читання текстів, у респондентів різних вікових груп.

В таблиці вказані значення середніх для різних груп:

- 36,5 с для респондентів віком до 22 років;
- 39,3 с для респондентів віком від 23 до 26 років;
- 41,7 с для респондентів віком від 27 років.

Оскільки гіпотеза  $H_1$  доведена, необхідно відповісти на питання, які

вікові групи статистично розрізняються. Для цього застосуємо t-критерій Стьюдента (рис. 54).

Двухвыборочный t-тест с различными дисперсиями			Двухвыборочный t-тест с различными дисперсиями			Двухвыборочный t-тест с различными дисперсиями		
	до 22 років	від 23 до 26 років		до 22 років	від 27 років		від 23 до 26 років	від 27 років
Среднее	36,46153846	39,27272727	Среднее	36,46153846	41,7	Среднее	39,27272727	41,7
Дисперсия	7,435897436	2,818181818	Дисперсия	7,435897436	47,34444444	Дисперсия	2,818181818	47,34444444
Наблюдения	13	11	Наблюдения	13	10	Наблюдения	11	10
Гипотетическая разность средних	0		Гипотетическая разность средних	0		Гипотетическая разность средних	0	
df	20		df	11		df	10	
t-статистика	-3,089049762		t-статистика	-2,274061845		t-статистика	-1,086526525	
P(T<=t) одностороннее	0,002893373		P(T<=t) одностороннее	0,021996326		P(T<=t) одностороннее	0,151370957	
t критическое одностороннее	1,724718243		t критическое одностороннее	1,795884819		t критическое одностороннее	1,812461123	
P(T<=t) двухстороннее	0,005786745		P(T<=t) двухстороннее	0,043992653		P(T<=t) двухстороннее	0,302741914	
t критическое двухстороннее	2,085963447		t критическое двухстороннее	2,20098516		t критическое двухстороннее	2,228138852	

Рис. 54. Результати обчислення Двохвибіркового t-тест з різними дисперсіями

Використання t-критерію Стьюдента дозволило встановити, що респонденти вікової групи до 22 років читають швидше, ніж респонденти вікових груп від 23 до 26 років та від 27 років (відмінності достовірні на рівні значущості  $p < 0,05$ ).

## 2. Практична робота із виконання дисперсійного аналізу в програмі Statistica.

Приклад. (Практична робота №9, Приклад 1.)

Рішення

1. Для використання інструменту **Breakdown & one-way ANOVA** вхідні дані на аркуші Statistica повинен бути представлений так, як показано на рисунку 55.

Data: Spreadsheet* (10v by 40c)			
	1	2	3
	Var1	Var2	Var3
1	35 до 22 років		
2	37 до 22 років		
3	43 до 22 років		
4	35 до 22 років		
5	35 до 22 років		
6	39 до 22 років		
7	40 до 22 років		
8	34 до 22 років		
9	34 до 22 років		
10	34 до 22 років		
11	35 до 22 років		
12	36 до 22 років		
13	37 до 22 років		
14	41 від 23 до 26 років		
15	40 від 23 до 26 років		
16	42 від 23 до 26 років		
17	39 від 23 до 26 років		
18	38 від 23 до 26 років		
19	37 від 23 до 26 років		
20	37 від 23 до 26 років		
21	38 від 23 до 26 років		
22	40 від 23 до 26 років		
23	41 від 23 до 26 років		
24	39 від 23 до 26 років		
25	46 від 27 років		
26	40 від 27 років		
27	40 від 27 років		
28	39 від 27 років		
29	38 від 27 років		
30	38 від 27 років		

Рис. 55. Таблиця з вхідними даними

Рішення

1. Відкрийте стартове вікно модуля **Basic Statistics/Tables – Основні статистики/Таблиці**.



2. Клацніть у стартовому вікні модуля на рядку **Breakdown & one-way ANOVA – Розбивка та односторонній дисперсійний аналіз**. Наступне вікно з'явиться на моніторі (рис. 56).

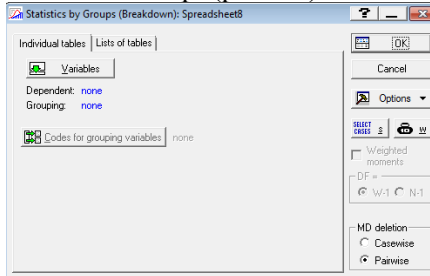


Рис. 56. Початкове вікно діалогу Однофакторного дисперсійного аналізу

3. Натисканням кнопки **Variables – Змінні** розкрийте вікно вибору змінних. У лівому списку (**Залежна змінна**) виберіть змінну (**Var1**), у правому (**Групуюча**) – (**Var2**). Далі клацніть **OK** або просто натисніть **Enter** на клавіатурі.

4. Потім натисніть кнопку **Codes for grouping variables – Коди для групування змінних** та натисніть **All – Всі**. Далі клацніть два рази **OK**.

Далі система здійснить аналіз заданого плану та виведе отримані результати у вікні **Statistics by Groups - Results: Spreadsheet – Статистика по групах - Результати: Таблиця**. За допомогою цього вікна користувач може звернутися до різних описових графіків та таблиць результатів (рис. 57).

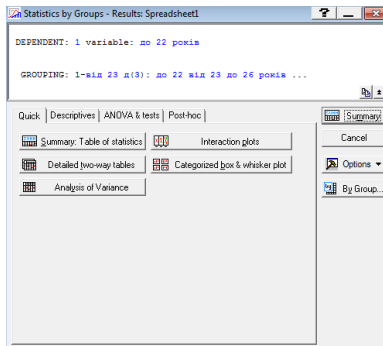


Рис. 57. Вікно *Statistics by Groups - Results: Spreadsheet – Статистика по групах - Результати: Таблиця*

5. На вкладці **Quick – Швидко** виберіть **Summary: Table of statistics - Короткий зміст: Таблиця статистики** та натисніть **Summary**. Отримаємо таблицю з результатом (рис. 58).

Breakdown Table of Descriptive Statistics (Spreadsheet8)				
N=34 (No missing data in dep. var. list)				
Var2	Var1		Var1	
	Means	N	Std.Dev.	
до 22 років	36,46154	13	2,726884	
від 23 до 26 років	39,27273	11	1,678744	
від 27 років	41,70000	10	6,880730	
All Grps	38,91176	34	4,608080	

Рис. 58. Breakdown Table of Descriptive Statistics - Таблиця розбивки описової статистики

В даній таблиці представлені результати середнього арифметичного значення, середнього квадратичного відхилення та кількість показників в кожній групі.

6. Знову відкрійте вікно **Statistics by Groups - Results: Spreadsheet – Статистика по групах - Результати: Таблиця** натиснувши клавіші Ctrl+R та натисніть кнопку **Analysis of Variance – Дисперсійний аналіз**. В результаті отримуємо таблицю результатів аналізу (Рядок Var1) (рис. 59). Параметри, на які потрібно звернути увагу: критерій Фішера  $F$  і рівень значущості  $p$ .

Analysis of Variance (Spreadsheet8)								
Marked effects are significant at $p < ,05000$								
Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Var1	157,2227	2	78,61135	543,5126	31	17,53266	4,483708	0,019483

Рис. 59. Підсумкова таблиця дисперсійного аналізу

У розглянутому прикладі **F-критерій** показує, що відмінність між групами є статистично значима на рівні  $p < 0,05$ .

Оскільки відмінності між середніми значеннями мають значення, нульова гіпотеза відкидається і приймається альтернативна гіпотеза про значущість відмінностей між групами (результат у рядку Var1 підсвічується **червоним кольором**).

**Висновок:** Оскільки розрахункове значення **критерію Фішера  $F$**  більше його **критичного значення на рівні значимості (альфа) 0,05**, то нульова гіпотеза відкидається. Рівень значущості  $p$  менше рівня значущості. Таким чином, можна стверджувати, що середні принаймні двох вибірок статистично значуще розрізняються.

7. Знову відкрійте вікно **Statistics by Groups - Results: Spreadsheet – Статистика по групах - Результати: Таблиця** натиснувши клавіші Ctrl+R та натисніть кнопку **Categorized box and whisker plot – Категоризований графік**. Потрібний графік буде відображено на екрані (рис. 60).

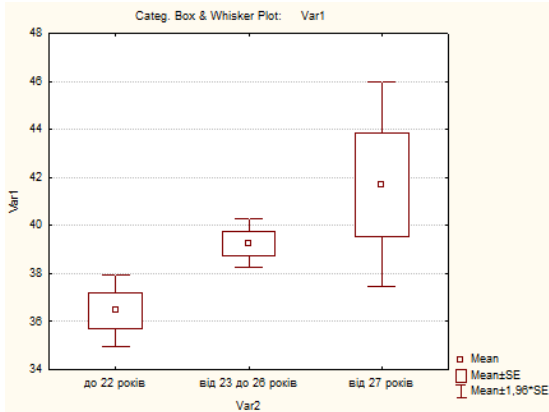


Рис. 60. Графічна інтерпретація результату дослідження

Оскільки гіпотеза  $H_1$  доведена, необхідно відповісти на питання, які вікові групи статистично розрізняються. Для цього застосуємо t-критерій Стьюдента **t-test, independent, by variables** - t-критерій для незалежних вибірок (рис. 61).

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
до 22 років vs. від 23 до 26 років		36,46154	39,27273	-2,97034	22	0,007063	13	11	2,726884	1,678744	2,638544	0,134029

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
до 22 років vs. від 27 років		36,46154	41,70000	-2,51407	21	0,020155	13	10	2,726884	6,880730	6,367011	0,004168

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
від 23 до 26 років vs. від 27 років		39,27273	41,70000	-1,13611	19	0,270039	11	10	1,678744	6,880730	16,79964	0,000130

Рис. 61. Результати обчислення T-критерію Стьюдента

Використання t-критерію Стьюдента дозволило встановити, що респонденти вікової групи до 22 років читають швидше, ніж респонденти вікових груп від 23 до 26 років та від 27 років (відмінності достовірні на рівні значущості  $p < 0,05$ ).

### Практична частина.

1. В таблиці представлені дані про середню успішність 6 здобувачів вищої освіти з 1-го по 5-й курс (відібраних випадковим чином) за підсумками двох сесій. Чи можна стверджувати, що існує тенденція

зростання середньої успішності при переході здобувачів вищої освіти з курсу на курс?

№	Курси				
	1	2	3	4	5
1	3,0	3,2	3,0	3,6	3,8
2	3,2	3,5	3,6	4,0	4,0
3	3,5	4,0	4,2	4,2	4,8
4	3,6	4,5	4,5	4,5	5,0
5	4,2	4,6	4,8	5,0	5,0
6	5,0	4,8	5,0	5,0	5,0

2. Хлопчики, прийняті у ДЮСШ на відділення спортивної гімнастики, протягом місяця проходили випробування у контрольній вправі – підтягуванні на перекладині. Усі три випробування було проведено у початковий період акцентованої загальної фізичної підготовки. Оцініть прогрес спортсменів у рівні розвитку сили.

№ досліджуваного	Кількість підтягувань		
	1 випробування	2 випробування	3 випробування
1	4	7	14
2	7	10	16
3	6	10	17
4	5	8	12
5	8	12	15
6	5	9	13
7	8	11	13
8	6	11	14

## Практична робота № 10

### Тема «Регресійний аналіз у програмі MS Excel»

**Мета заняття:** ознайомитися із методом регресійного аналізу та способами розрахунку в програмі MS Excel

#### Зміст і хід заняття

1. Теоретичні основи регресійного аналізу
2. Обробка даних у межах лінійної регресійної моделі
3. Аналіз взаємозв'язків багатомірних даних (множинна регресія).

Виконання та захист практичного завдання.

#### 1. Теоретичні основи регресійного аналізу

Регресія. Категорія завдань, де мета полягає в тому, щоб оцінити значення безперервної вихідної змінної за значеннями вхідних змінних.

У найпростішому випадку в лінійній регресійній моделі є дві змінні  $X$  і  $Y$ . І потрібно по парам спостережень  $(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$  побудувати

(підібрати) пряму лінію, яка "найкращим чином" наближає значення, що спостерігаються.

Загальне призначення множинної регресії полягає в аналізі зв'язку між декількома незалежними змінними (названими також регресорами) і залежною змінною.

Іноді використовується інша термінологія. Залежна змінна часто називається відгуком, а незалежні змінні – предикторами, контрольовані змінні – факторами. Ця термінологія підкреслює, що ряд змінних впливає на одну змінну – відгук.

Загальна обчислювальна задача, яку потрібно вирішувати при аналізі методом множинної регресії, полягає у припасуванні прямої лінії до деякого набору точок.

Рівняння регресії. Пряма лінія на площині (у просторі двох вимірів) задається рівнянням

$$\hat{y} = a + b \times x$$

докладніше: змінна  $y$  може бути виражена через константу ( $a$ ) і кутовий коефіцієнт ( $b$ ), помножений на змінну  $x$ . Константу іноді називають також вільним членом, а кутовий коефіцієнт – регресійним або  $\beta$ -коефіцієнтом.

*Метод найменших квадратів.* На діаграмі розсіювання є незалежна змінна або змінна  $x$  і залежна змінна  $y$ . Кожна точка на діаграмі представляє дані одного спостереження. Метою процедур лінійної регресії є припасування прямої лінії по точках. А саме, програма будує лінію регресії так, щоб мінімізувати квадрати відхилень цієї лінії від точок, що спостерігаються. Тому на цю загальну процедуру іноді посилаються як на оцінювання методом найменших квадратів.

*Передбачені значення та залишки.* Лінія регресії відображає найкраще передбачення залежної змінної ( $y$ ) по незалежним змінним ( $x$ ). Однак, природа рідко (якщо взагалі коли-небудь) буває цілком передбачуваною і зазвичай є суттєвий розкид точок щодо підігнаної прямої (як це було показано раніше на діаграмі розсіювання). Відхилення окремої точки від лінії регресії (від передбаченого значення) називається залишком.

*Залишкова дисперсія та коефіцієнт детермінації R-квадрат.* Чим менший розкид значень залишків біля лінії регресії стосовно загального розкиду значень, тим, очевидно, кращий за прогноз. Наприклад, якщо зв'язок між змінними  $x$  і  $y$  відсутній, то відношення залишкової мінливості змінної  $y$  до вихідної дисперсії дорівнює 1,0. Якщо  $x$  і  $y$  жорстко пов'язані, залишкова мінливість відсутня, і відношення дисперсій буде дорівнює 0,0. Найчастіше ставлення лежатиме десь між цими екстремальними значеннями, тобто між 0,0 та 1,0. 1,0 мінус це відношення називається R-квадратом або коефіцієнтом детермінації. Це значення безпосередньо

інтерпретується в такий спосіб. Якщо  $R$ -квадрат дорівнює 0,4, то мінливість значень змінної  $y$  біля лінії регресії становить 1-0,4 від вихідної дисперсії; іншими словами, 40% від вихідної мінливості можуть бути пояснені, а 60% залишкової мінливості залишаються непоясненими. В ідеалі бажано мати пояснення якщо не для всієї, то хоча б для більшої частини вихідної мінливості. Значення  $R$ -квадрата є індикатором ступеня підгонки моделі до даних (значення  $R$ -квадрата близьке до 1,0 показує, що модель пояснює майже всю мінливість відповідних змінних).

*Інтерпретація коефіцієнта множинної кореляції  $R$ .* Зазвичай ступінь залежності двох або більше предикторів (незалежних змінних або змінних  $X$ ) із залежною змінною ( $y$ ) виражається за допомогою коефіцієнта множинної кореляції  $R$ . За визначенням він дорівнює квадратному кореню з коефіцієнта детермінації. Це невід'ємна величина, що приймає значення між 0 і 1. Для інтерпретації спрямування зв'язку між змінними дивляться на знаки (плюс або мінус) регресійних коефіцієнтів або  $\beta$ -коефіцієнтів. Якщо  $\beta$ -коефіцієнт позитивний, то зв'язок цієї змінної із залежною змінною позитивна; якщо  $\beta$ -коефіцієнт негативний, то й зв'язок носить негативний характер. Звичайно, якщо  $\beta$ -коефіцієнт дорівнює 0, зв'язок між змінними відсутній.

## 2. Обробка даних у межах лінійної регресійної моделі

**Приклад 1.** Необхідно оцінити залежність результату у велоперегонах на дистанції 200 м ( $y_i$ , с) від часу педалювання 30 обертів на велоергометрі ( $x_i$ , с). Дані вимірювань представлені на рисунку 62.

	A	B
1	$x_j$	$y_j$
2	11,0	18,3
3	10,7	18,1
4	10,5	17,9
5	10,3	17,5
6	10,2	17,2
7	10,0	17,1

Рис. 62. Таблиця з вхідними даними

Вирішимо цю задачу з використанням програми **Microsoft Excel**.

Створити таблицю з числовими даними за зразком.

Побудувати точкову діаграму: виділити діапазон із вхідними даними 2-х стовпців – обрати меню **Вставка – Точкова діаграма** (рис. 63).

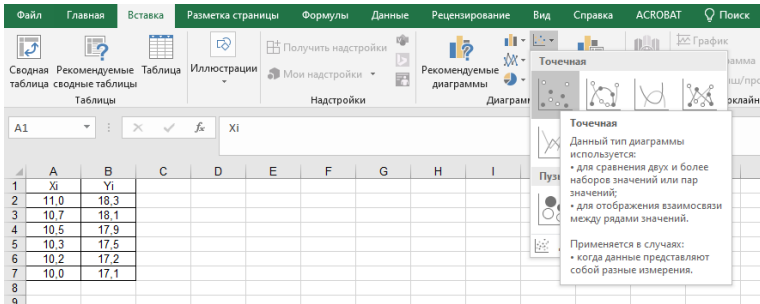


Рис. 63. Побудова точкової діаграми

Діаграма матиме наступний вигляд (рис. 64).

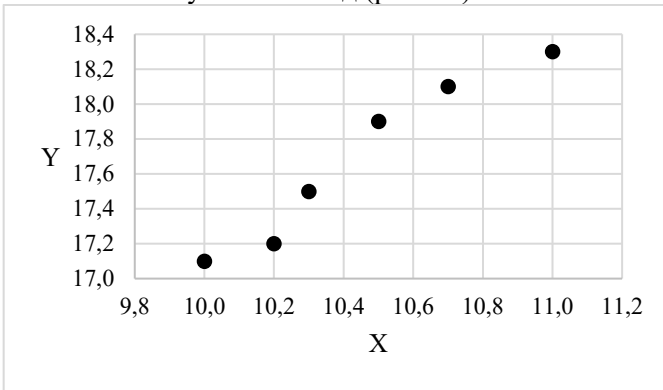


Рис. 64. Точкова діаграма

За розташуванням точок на графіку можна зробити висновок, що між даними ознаками існує взаємозв'язок позитивного характеру.

Додати лінію тренду на графік: натиснути правою кнопкою миші на будь-яку точку діаграми – обрати пункт **Додати лінію тренду** (рис. 65).

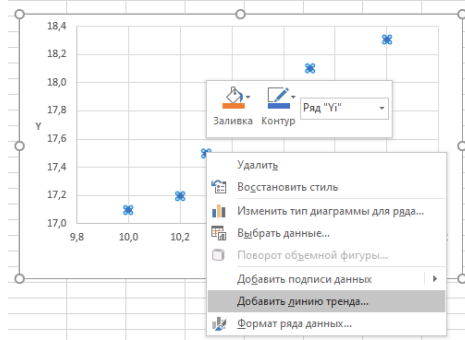


Рис. 65. Додавання лінії тренду

Результат представлено на рис. 66.

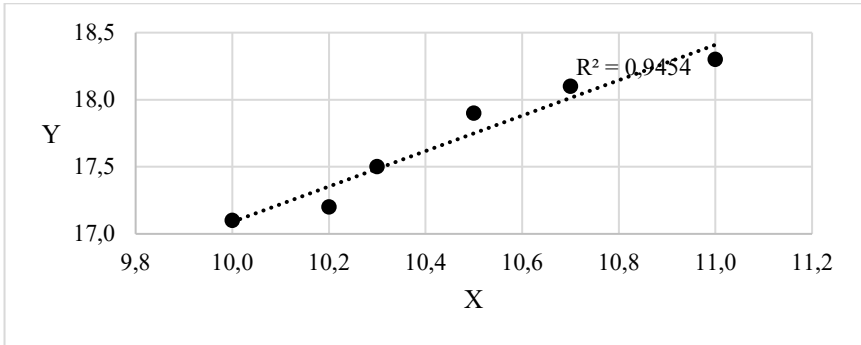


Рис. 66. Лінія тренду на діаграмі (вид лінії – лінійна)

За результатом побудови лінії тренду та аналізу коефіцієнту детермінації ( $R=0,9454$ ) можна зробити висновок, що залежність між даними ознаками можна описати рівнянням лінійної регресії:

$$\hat{y} = b_0 + bx$$

Визначимо параметри рівняння регресії та адекватність даної регресійної моделі. У горизонтальному меню **Дані** вибрати пункт **Аналіз даних – Регресія** (рис. 67).

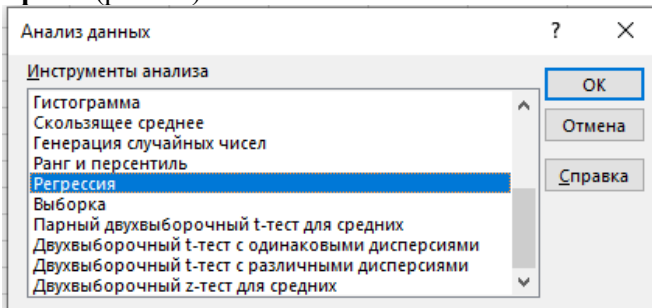


Рис. 67. Діалогове вікно вибору статистичного методу у меню **Аналіз даних**

У полі **Вхідний інтервал Y** вибрати результати у велоерегонах на дистанції 200 м (залежна змінна  $Y$ ).

У полі **Вхідний інтервал X** вибрати час педалювання (с) 30 обертів на велоергометрі (незалежна змінна  $X$ ).

Встановити прапорець у полях **Мітки** та **Рівень надійності** (95%).

Вибрати **Вихідний інтервал** (будь-який осередок) (рис. 68).



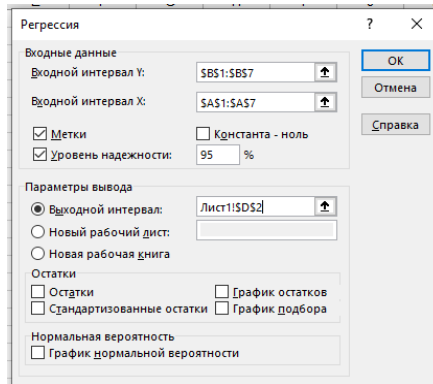


Рис. 68. Диалогове вікно регресійного аналізу

Оцінюємо коефіцієнт детермінації (***R-квадрат***). Оскільки значення коефіцієнту детермінації близький до 1, можна зробити висновок, що мінливість залежної змінної  $y$  на 95% обумовлена значеннями незалежної змінної  $x$ .

Перевіряємо статистичну значущість коефіцієнту детермінації за критерієм Фішера (***F, Значущість F***). Значущість за критерієм Фішера знаходиться на рівні  $p < 0,01$ , що свідчить про статистичну значущість коефіцієнту детермінації (адекватність регресійної моделі) (рис. 69).

	D	E	F	G	H	I	J	K	L
Вывод ИТОГОВ									
<b>Регрессионная статистика</b>									
Множественный R		0,972304016							
R-квадрат		0,945375099							
Нормированный R-квадрат		0,931718873							
Стандартная ошибка		0,128457278							
Наблюдения		6							
<b>Дисперсионный анализ</b>									
		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>			
Регрессия		1	1,142328244	1,142328244	69,22667695	0,001139979			
Остаток		4	0,066005089	0,016501272					
Итого		5	1,208333333						
		<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>	<i>Нижние 95,0%</i>	<i>Верхние 95,0%</i>
Y-пересечение		3,882951654	1,659477327	2,33886424	0,079390147	-0,724496047	8,490399355	-0,724496047	8,490399355
X1		1,320610687	0,158722343	8,320257024	0,001139979	0,879926815	1,761294559	0,879926815	1,761294559

Рис. 69. Результат регресійного аналізу

Перевіримо значущість **коефіцієнту незалежної змінної  $x$**  рівняння регресії. Коефіцієнт має рівень значущості  $p < 0,01$ , тому від може бути включений до рівняння регресії.

Значення **коефіцієнту  $b_0$**  знаходиться у рядку **Y-перетин**, значення **коефіцієнту незалежної змінної  $x$**  у рядку  **$x_1$** .

Складаємо рівняння регресії:

$$\hat{y} = 3,88 + 1,32x$$

Дане рівняння дозволяє прогнозувати можливе значення залежної змінної (результати у велоперегонах на дистанції 200 м) від значень незалежної змінної (часу педалювання (с) 30 обертів на велоергометрі).

### 3. Аналіз взаємозв'язків багатомірних даних (множинна регресія).

Багатовимірний регресійний аналіз визначає:

- значення однієї залежної (результативної) змінної з множини значень інших незалежних змінних;
- міру і статистичну значимість взаємозв'язку результативної ознаки за сукупністю факторних ознак;
- суттєвість вкладу кожної факторної ознаки в оцінку результативної ознаки;
- точність передбачення і ймовірності помилок оцінки результативної ознаки;
- невідомі значення результативної ознаки за рівнянням регресії.

Треба враховувати, що всі змінні повинні бути виміряні в кількісній шкалі.

Одна з змінних визначається дослідником як залежна ( $Y$ ), решта (або їх частина!) – як незалежні змінні.

При цьому передбачається, що зв'язок між однією залежною змінною ( $Y$ ) і декількома незалежними змінними ( $X$ ) можна виразити лінійним рівнянням:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + \varepsilon$$

яке включає:  $b_0$  – вільний член (Intercept), коефіцієнти регресії –  $b_1, \dots, b_k$ , (Unstandardized Coefficients),  $\varepsilon$  – помилка передбачень (Residual).

Приклади:

$Y$  – успішність студента,  $X_1 \dots X_k$  – когнітивні якості;

$Y$  – оплата праці,  $X_1 \dots X_k$  – вік, стаж і т. д.;

$Y$  – спортивний результат,  $X_1 \dots X_k$  – фізичні якості.

Розглянемо приклад виконання множинного регресійного аналізу.

**Приклад 2.** В групі спортсменів проведено визначення наступних показників: індекс СОЗ (бал), індекс маси тіла ( $\text{кг}/\text{м}^2$ ), життєвий індекс ( $\text{мл}/\text{кг}$ ), індекс Робінсона, силовий індекс (%), час відновлення (с), а також розраховано біологічний вік. Необхідно встановити регресійну залежність між біологічним віком та показниками індексів. Визначити, які показники впливають на біологічний вік (за рівнем взаємозв'язку).

Вирішимо цю задачу з використанням програми **Microsoft Excel**. Результати виміру представлені у таблиці 11.

Таблиця 11

Вхідні дані

Індекс СО <sub>2</sub> , бал	Індекс маси тіла, кг/м <sup>2</sup>	Життєвий індекс, мл/кг	Індекс Робінсона	Силовий індекс, %	Час відновлення, с	Біологічний вік
1	20,66	64,41	112,32	27,12	90,00	24,99
5	21,38	52,17	87,36	31,30	80,00	31,52
7	20,99	66,17	85,80	39,10	120,00	21,29
2	17,34	72,98	72,00	27,61	90,00	21,34
8	17,51	54,72	74,00	15,09	80,00	32,25
7	21,71	47,58	90,20	16,13	70,00	33,10
2	19,85	54,31	76,80	22,47	90,00	30,49
3	19,71	49,73	84,60	26,64	90,00	24,29
9	20,55	43,10	85,50	22,41	120,00	26,76
5	20,94	42,11	118,08	33,33	120,00	40,11
5	22,76	56,92	101,68	23,08	90,00	26,24
5	23,99	55,09	96,20	23,37	80,00	27,83
6	27,70	50,14	98,40	33,43	60,00	18,40
7	22,50	50,80	95,94	27,58	150,00	33,90
5	18,18	55,21	97,44	26,58	59,00	35,17
4	20,20	57,89	78,00	28,07	80,00	28,30
3	19,84	63,79	79,20	13,79	140,00	22,60
16	22,05	45,37	90,00	26,02	80,00	40,44
10	21,80	58,91	72,00	31,01	59,00	27,46
6	21,99	55,06	97,75	26,64	150,00	21,07
8	22,99	43,13	87,36	19,17	150,00	33,36
3	19,91	69,77	83,22	24,81	90,00	15,18
8	22,43	47,62	73,00	42,33	80,00	33,25
2	20,58	57,41	53,69	35,19	59,00	21,67
9	20,18	57,80	87,15	28,81	150,00	23,81

1	2	3	4	5	6	7
2	21,72	64,62	121,03	27,69	90,00	20,55
11	22,31	58,33	134,42	41,67	120,00	34,30
14	21,19	47,13	76,00	32,40	90,00	24,25
5	22,31	47,73	102,70	28,79	120,00	33,18
7	20,42	50,00	72,00	20,34	90,00	27,91
7	20,81	55,18	89,68	31,77	60,00	29,23
10	22,18	55,28	95,68	19,51	150,00	32,43
7	19,72	63,64	103,88	40,00	90,00	29,07
9	18,67	62,37	83,60	33,26	59,00	33,30
6	22,13	59,52	111,72	30,61	60,00	32,02
1	20,96	64,15	81,00	24,53	60,00	23,33
9	20,05	53,33	91,20	30,00	90,00	32,02
1	22,91	50,75	64,80	38,81	60,00	27,07
4	21,98	67,37	105,60	21,05	80,00	33,57
3	19,44	52,58	96,00	39,68	90,00	29,68
10	18,77	64,37	79,20	27,99	59,00	34,08
5	18,78	57,62	77,76	29,70	105,00	29,25
3	18,15	44,15	48,72	30,91	59,00	28,32
12	21,07	56,06	85,60	19,70	180,00	33,21

Перейдіть до меню **Дані** → **Сервіс** → **Аналіз даних** → **Регресія**.

У полі **Вхідний інтервал Y** обрати біологічний вік (залежна змінна).

В поле **Вхідний інтервал X** обрати показники індексів (незалежні змінні).

Встановити прапорець у полі **Рівень надійності (95%)**, **Мітки**, **Константа нуль**.

Вибрати **Вихідний інтервал** (будь-який осередок на аркуші з вихідними даними) (рис. 70).

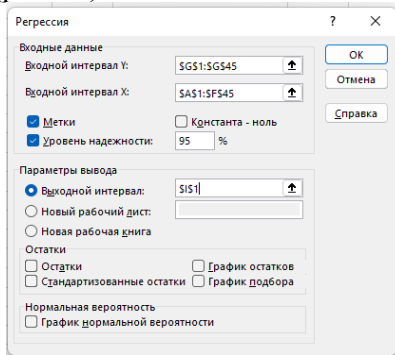


Рис. 70. Діалогове вікно регресійного аналізу

Аналізуємо дані третьої (нижньої) таблиці результатів. Кожна незалежна змінна має певний рівень значущості, що знаходиться у 5 стовпці останньої таблиці (***P*-значення**) – осередки N18:N24. Для включення у рівняння регресії обираємо ті змінні, які мають достовірний рівень значущості ( $p < 0,05$ ). В даному прикладі це змінні індекс СО<sub>2</sub> ( $p < 0,05$ ), індекс маси тіла ( $p < 0,05$ ), життєвий індекс ( $p < 0,01$ ) та індекс Робінсона ( $p < 0,01$ ) (рис. 71).

	J	K	L	M	N	O	P	Q	R
<b>ВЫВОД ИТОГОВ</b>									
<i>Регрессионная статистика</i>									
Множественный R			0,701643296						
R-квадрат			0,492303314						
Нормированный R-квадрат			0,409974122						
Стандартная ошибка			4,282208789						
Наблюдения			44						
<i>Дисперсионный анализ</i>									
		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>			
Регрессия		6	657,9090077	109,6515013	5,979693239	0,000192599			
Остаток		37	678,4805482	18,33731211					
Итого		43	1336,389556						
		<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>	<i>Нижние 95,0%</i>	<i>Верхние 95,0%</i>
У-пересечение		60,10375769	11,1413783	5,394642933	4,14561E-06	37,52918096	82,67833442	37,52918096	82,67833442
Индекс СО <sub>2</sub> , бал		0,568208618	0,206158458	2,756174178	0,009024331	-0,150491905	0,985925332	-0,150491905	0,985925332
Индекс маси тіла, кг/м <sup>2</sup>		-1,096922701	0,403059987	-2,721487462	0,009848358	-1,913599809	-0,280245593	-1,913599809	-0,280245593
Життєвий індекс, мл/кг		-0,368350789	0,098392418	-3,743690813	0,000615199	-0,567712764	-0,168988815	-0,567712764	-0,168988815
Индекс Робинсона		0,140544809	0,045427343	3,093837349	0,003750487	0,04850027	0,232589348	0,04850027	0,232589348
Силовий індекс, %		-0,040959911	0,098658634	-0,415168029	0,680416624	-0,240861292	0,158941471	-0,240861292	0,158941471
t відновлення, с		-0,030127148	0,022693011	-1,327595894	0,192444882	-0,076107555	0,015853259	-0,076107555	0,015853259

Рис. 71. Результат регресійного аналізу з 6 змінними

На наступному кроці повертаємося до меню **Дані** → **Сервіс** → **Аналіз даних** → **Регресія** і обираємо в якості незалежних змінних X тільки ті змінні, що мають достовірний рівень значущості (рис. 72).

J	K	L	M	N	O	P	Q	R
<b>Вывод ИТОВОВ</b>								
<i>Регрессионная статистика</i>								
Множественный R	0,684187116							
R-квадрат	0,468112009							
Нормированный R-квадрат	0,413559395							
Стандартная ошибка	4,269178614							
Наблюдения	44							
<i>Дисперсионный анализ</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>			
Регрессия	4	625,5800004	156,3950001	8,580927137	4,56063E-05			
Остаток	39	710,8095555	18,22588604					
Итого	43	1336,389556						
	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>	<i>Нижние 95,0%</i>	<i>Верхние 95,0%</i>
Y-пересечение	56,89956738	10,67942956	5,327959426	4,42908E-06	35,29838218	78,50075257	35,29838218	78,50075257
Индекс СО2, бал	0,527325655	0,202751679	2,600844818	0,013071142	0,117221674	0,937429636	0,117221674	0,937429636
Индекс маси тіла, кг/м2	-1,088197746	0,400703156	-2,715720429	0,009804468	-1,89869638	-0,277699111	-1,89869638	-0,277699111
Життєвий індекс, мл/кг	-0,355248644	0,097598833	-3,639904944	0,000790198	-0,552659901	-0,157837388	-0,552659901	-0,157837388
Индекс Робинсона	0,124299271	0,043559778	2,853533158	0,006886459	0,036191304	0,212407239	0,036191304	0,212407239

Рис. 72. Результат регресійного аналізу з 4 змінними

Проаналізуємо результати.

Коефіцієнт детермінації дорівнює 0,47 (осередок К33), що свідчить про відсоток поясненої дисперсії залежної змінної незалежними (факторними) змінними – 47%.

Перевірка коефіцієнта детермінації за критерієм Фішера у ході дисперсійного аналізу показала, що він має високий рівень значущості  $p < 0,001$  – осередок О40. Цей факт підтверджує адекватність даної регресійної моделі.

Переходимо до складання рівняння регресії. У нижній таблиці результатів у стовпці **Коефіцієнти** вказані значення  $b_0$  та параметри рівняння регресії для незалежних змінних. Значення  $b_0$ , знаходиться як **Y-перетин**, нижче у таблиці – значення незалежних змінних **X1 – X4**. Для даних значень рівняння регресії має вигляд:

$$\hat{y} = 56,90 + 0,53x_1 - 1,09x_2 - 0,36x_3 + 0,12x_4$$

Дане рівняння являє собою математичну модель регресійної залежності між показниками біологічного віку та індексу СО2 ( $x_1$ ), індексу маси тіла ( $x_2$ ), життєвого індексу ( $x_3$ ) та індекс Робінсона ( $x_4$ ).

### Практична частина.

1. Визначте взаємозв'язок довжини кроку в бігу (см) від величини споживання кисню (л·хв.<sup>-1</sup>). Зробіть висновок.

1) Визначити чисельну міру залежності для емпіричних даних (кореляційний аналіз).

2) Побудуйте точкову діаграму для графічного зображення існуючого взаємозв'язку, додайте лінію тренду.

3) Складіть рівняння регресії (споживання кисню – залежна змінна).

Довжина кроку, см	Споживання кисню, л·хв. <sup>-1</sup>
132	4,1
138	4
143	3,9
146	3,9
150	4
154	4,1
159	4,2
163	4,2
170	4,6

2. У регбістів вимірювали силу кидка м'яча,  $x_i$  (Н) та дальність польоту м'яча,  $y_i$  (м).

1) Визначити чисельну міру залежності для емпіричних даних (кореляційний аналіз).

2) Складіть рівняння регресії (дальність польоту м'яча – залежна змінна).

№	Сила кидка м'яча, Н	Дальність польоту м'яча, м	№	Сила кидка м'яча, Н	Дальність польоту м'яча, м
1	10,12	25,21	7	10,14	25,23
2	10,31	26,42	8	10,32	26,43
3	10,65	27,24	9	10,67	27,22
4	11,21	27,92	10	11,02	27,93
5	11,92	28,51	11	11,92	28,52
6	12,31	31,20	12	12,32	31,23

3. На змаганнях з баскетболу досліджували точність влучення м'яча до кошика. Фіксувалися два показники: швидкість м'яча та кут вильоту м'яча, що потрапив у кошик.

1) Визначити чисельну міру залежності для емпіричних даних (кореляційний аналіз).

2) Складіть рівняння регресії (кут вильоту – залежна змінна).

№	Швидкість, м/с	Кут вильоту	№	Швидкість, м/с	Кут вильоту
1	27	5,9	9	42	5,1
2	27	5,7	10	45	4,7
3	30	5,6	11	45	4,9
4	33	5,5	12	48	5
5	33	5,3	13	51	5
6	36	5,3	14	54	5,2
7	39	5,2	15	57	5,3
8	42	4,9			

4. Вивчити залежність життєвої ємності легень (ЖЄЛ) у літрах ( $y$ ) від зросту в метрах ( $x_1$ ) та віку у роках ( $x_2$ ) для групи чоловіків.

1) Визначити чисельну міру залежності для емпіричних даних (кореляційний аналіз).

2) Складіть рівняння регресії (життєвої ємності легень – залежна змінна).

Зріст, м	Вік, років	ЖЄЛ, л
1,85	18	5,4
1,80	25	6,5
1,75	20	4,8
1,70	24	5,1
1,68	21	4,5
1,73	19	4,8
1,77	22	5,1
1,81	23	5,6
1,76	18	4,7

5. Вивчити залежність часу за сумою двох заїздів (скелетон) із результатами спеціальних вправ.

1) Визначити чисельну міру залежності для емпіричних даних (кореляційний аналіз).

2) Складіть рівняння регресії (сумою двох заїздів – залежна змінна).

Сума двох заїздів (с)	Потрійний стрибок у довжину з місця (см)	Метання ядра спиною 4 кг (см)	Стартовий розгін (с)	10 м біг з візком 20 кг (с)
116,99	811	1333	4,92	1,62
120,92	681	964	5,27	1,89
119,11	667	942	5,46	1,96
119,80	709	1085	5,2	1,80
117,50	762	1196	5,11	1,74
117,01	743	1090	5,08	1,72
120,95	701	1049	5,31	1,84
118,30	720	1036	5,22	1,76
120,83	729	1110	5,29	1,88
119,39	705	1125	5,35	1,79

## Практична робота № 11

### Тема «Регресійний аналіз у програмі Statistica»

**Мета заняття:** ознайомитися із методом регресійного аналізу та способами розрахунку в програмі Statistica

#### Зміст і хід заняття

1. Практична робота із розрахунку регресійного аналізу в програмі Statistica.

Виконання та захист практичного завдання.



## 1. Практична робота із розрахунку регресійного аналізу в програмі Statistica.

### Парний лінійний регресійний аналіз

**Приклад 1.** Необхідно оцінити залежність результату у велоперегонах на дистанції 200 м ( $y_i$ , с) від часу педалювання 30 обертів на велоергометрі ( $x_i$ , с). Дані вимірювань представлені на рисунку 73.

	1	2
	$X_i$	$Y_i$
1	11,0	18,3
2	10,7	18,1
3	10,5	17,9
4	10,3	17,5
5	10,2	17,2
6	10,0	17,1

Рис. 73. Таблиця з вхідними даними

Вирішимо дану задачу з використанням програми Statistica. Створити новий файл. Для цього в меню **Файл** вибрати пункт **New**. У меню **Statistics (Статистики)** вибрати пункт **Multiple Regression (Множинна регресія)** (рис. 74).

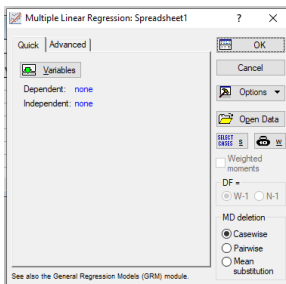


Рис. 74. Вибір модуля регресійного аналізу

У новому вікні натиснути кнопку **Variabes (Змінні)** та обрати залежну (**Dependent, Y**) і незалежну (**Independent, X**) змінні, після чого натиснути **OK** (рис. 75).

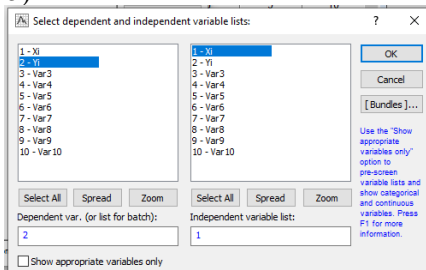


Рис. 75. Вибір змінних

**Аналіз результату.** Опишемо вікно результатів кореляційно-регресійного аналізу. 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. Під лінією подані стандартизовані оцінки коефіцієнтів регресії (рис. 76).

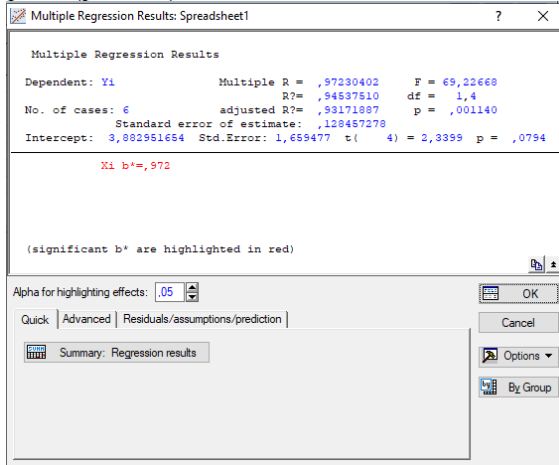


Рис. 76. Діалогове вікно регресійного аналізу

**Dependent: Yi** – ім'я залежної змінної (результати у велоперегонах на дистанції 200 м).

**No. of cases** – число спостережень, за якими побудована регресія (6).

**Multiple R** – коефіцієнт множинної кореляції (описує залежності між змінними).

**R-square** – коефіцієнт детермінації – показує частку загального розкиду щодо вибіркового середнього залежної змінної, яка пояснюється побудованою регресією.

**Adjusted R-square** – скоригований коефіцієнт детермінації:

$$\text{Adjusted R-square} = 1 - (1 - \text{R-square}) \cdot (n / (n - p)),$$

де  $n$  – число спостережень моделі,  $p$  – число параметрів моделі.

**Standard error of estimate:** стандартна помилка оцінки – є мірою розсіювання спостережуваних значень відносно регресійної прямої.

**Intercept:** оцінка вільного члена регресії – значення коефіцієнта  $b_0$  в рівнянні регресії.

**Std. Error:** стандартна помилка оцінки вільного члена – стандартна помилка коефіцієнта  $b_0$  в рівнянні регресії.

**F** – значення F-критерію.

**df and p** - число ступенів свободи і рівень значущості F-критерію.

**t and p** – значення **t-критерію** і **рівня p** (t-критерій використовується для перевірки гіпотези про рівність нулю вільного члена регресії).

Натиснути кнопку **Summary: Regression results**. 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. 77, 78).

Regression Summary for Dependent Variable: Yi (Spreadsheet1)						
R= ,97230402 R²= ,94537510 Adjusted R²= ,93171887						
F(1,4)=69,227 p<,00114 Std.Error of estimate: ,12846						
	$b^*$	Std.Err. of $b^*$	b	Std.Err. of b	t(4)	p-value
N=6						
<b>Intercept</b>			3,882952	1,659477	2,339864	0,079390
$X_i$	0,972304	0,116860	1,320611	0,158722	8,320257	0,001140

Рис. 76. Вікно результатів регресійного аналізу

Summary Statistics:	
Statistic	Value
<b>Multiple R</b>	0,972304016
Multiple R²	0,945375099
Adjusted R²	0,931718873
F(1,4)	69,2266769
p	0,00113997899
Std.Err. of Estimate	0,128457278

Рис. 78. Вікно результатів регресійного аналізу (підсумкова статистика)

За даними отриманої таблиці складемо рівняння регресії:

$$\hat{y} = 3,88 + 1,32x$$

Побудуємо графік залежності двох змінних: в меню **Графіки (Graphs)** обрати пункт **Графіки розсіювання (Scatterplot)**. У вікні побудови графіка необхідно вибрати залежні незалежні змінні: натиснути кнопку **Змінні (Variabeles)**, вибрати для  $X$  змінну  $X_i$ , для  $Y$  – змінну  $Y_i$  – натиснути ОК (рис. 79).

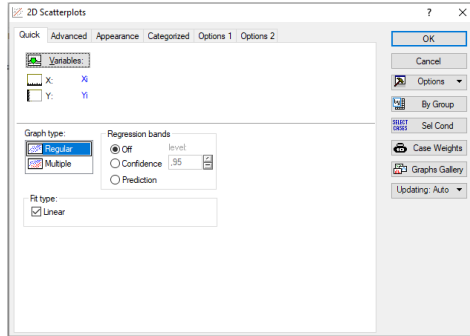


Рис. 79. Вікно побудови графіка розсіювання

На даному графіку видно лінію регресії і дві лінії залишків. У верхній частині поля виведено рівняння регресії (рис. 80).

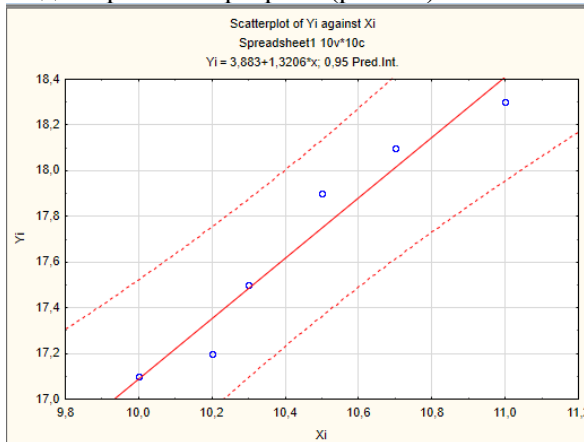


Рис. 80. Графік розсіювання

### Множинна регресія

**Приклад 2.** В групі спортсменів проведено визначення наступних показників: індекс СО<sub>2</sub> (бал), індекс маси тіла (кг/м<sup>2</sup>), життєвий індекс (мл/кг), індекс Робінсона, силовий індекс (%), час відновлення (с), а також розраховано біологічний вік. Необхідно встановити регресійну залежність між біологічним віком та показниками індексів. Визначити, які показники впливають на біологічний вік (за рівнем взаємозв'язку). Таблиця з вхідними даними представлена в Практичній роботі №10, Приклад 1.

Вирішимо цю задачу з використанням програми STatistica.

Створити новий файл. Для цього в меню **Файл** вибрати пункт **New** (рис. 81).

	1	2	3	4	5	6	7
	Індекс СО <sub>2</sub> , бал	Індекс маси тіла, кг/м <sup>2</sup>	Життєвий індекс, мл/кг	Індекс Робінсона	Силовий індекс, %	Час відновлення, с	Біологічний вік
1	5	20,66	64,41	112,32	27,12	90,00	24,99
2	7	21,38	52,17	87,36	31,30	80,00	31,52
3	2	20,99	66,17	85,80	39,10	120,00	21,29
4	4	17,34	72,88	72,00	27,61	90,00	21,34
5	8	17,51	54,72	74,00	15,09	80,00	32,25
6	7	21,71	47,58	90,20	16,13	70,00	33,10
7	2	19,85	54,31	76,80	22,47	90,00	30,49
8	3	19,71	49,73	84,60	26,64	90,00	24,29
9	9	20,55	43,10	85,50	22,41	120,00	26,76
10	5	20,94	42,11	118,08	33,33	120,00	40,11
11	5	22,76	56,92	101,68	23,08	90,00	26,24
12	5	23,99	55,09	96,20	23,37	80,00	27,83
13	6	27,70	50,14	98,40	33,43	60,00	18,40
14	7	22,50	50,80	95,94	27,58	150,00	33,90
15	5	18,18	55,21	97,44	26,58	59,00	35,17
16	4	20,20	57,89	78,00	28,07	80,00	28,30
17	3	19,84	63,79	79,20	13,79	140,00	22,60
18	16	22,05	45,37	90,00	26,02	80,00	40,44
19	10	21,80	58,91	72,00	31,01	59,00	27,46
20	6	21,99	55,06	97,75	26,64	150,00	21,07
21	8	22,99	43,13	87,36	19,17	150,00	33,36
22	3	19,91	69,77	83,22	24,81	90,00	15,18
23	8	22,43	47,62	73,00	42,33	80,00	33,25
24	2	20,58	57,41	53,69	35,19	59,00	21,67
25	9	20,18	57,80	87,15	28,81	150,00	23,81
26	2	21,72	64,62	121,03	27,69	90,00	20,55

Рис. 81. Таблиця з вихідними даними

У меню **Statistics (Статистики)** вибрати пункт **Multiple Regression (Множинна регресія)**.

У новому вікні натиснути кнопку **Variabes (Змінні)** та обрати залежну (**Dependent, Y**) і незалежні (**Independent, X**) змінні, після чого натиснути **OK** (рис. 82).

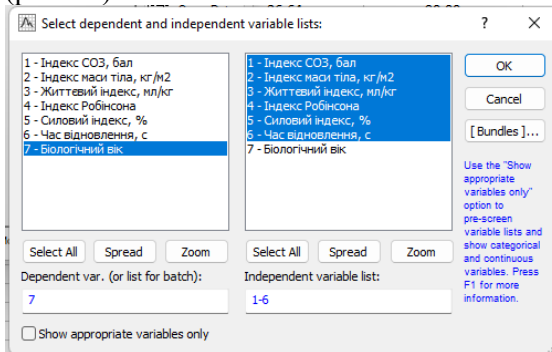


Рис. 82. Вибір змінних

**Аналіз результату.** Опишемо вікно результатів кореляційно-регресійного аналізу. 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. Під лінією подані стандартизовані оцінки коефіцієнтів регресії (рис. 83).

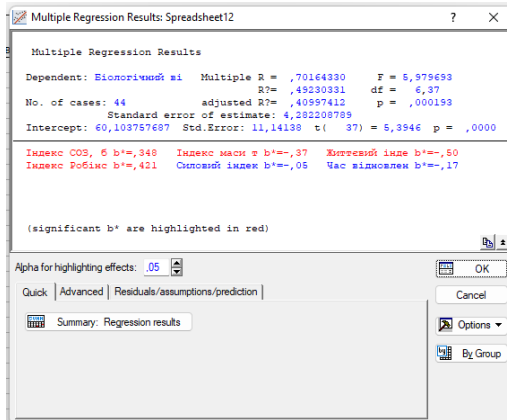


Рис. 83. Діалогове вікно регресійного аналізу

**Dependent: Yi** – ім'я залежної змінної (біологічний вік).

**No. of cases** – число спостережень, за якими побудована регресія (44).

**Multiple R** – коефіцієнт множинної кореляції (описує залежності між змінними).

**R-square** – коефіцієнт детермінації – показує частку загального розкиду щодо вибіркового середнього залежної змінної, яка пояснюється побудованою регресією.

**Adjusted R-square** – скоригований коефіцієнт детермінації:

$$\text{Adjusted R-square} = 1 - (1 - R\text{-square}) \cdot (n / (n - p)),$$

де  $n$  – число спостережень моделі,  $p$  – число параметрів моделі.

**Standard error of estimate:** стандартна помилка оцінки – є мірою розсіювання спостережуваних значень відносно регресійної прямої.

**Intercept:** оцінка вільного члена регресії – значення коефіцієнта  $b_0$  в рівнянні регресії.

**Std. Error:** стандартна помилка оцінки вільного члена – стандартна помилка коефіцієнта  $b_0$  в рівнянні регресії.

**F** – значення F-критерію.

**df and p** – число ступенів свободи і рівень значущості F-критерію.

**t and p** – значення **t-критерію** і **рівня p** (t-критерій використовується для перевірки гіпотези про рівність нулю вільного члена регресії).

Бачимо, що частина коефіцієнтів мають значущості на рівні  $p < 0,05$ . На цьому етапі приймаємо рішення про включення до рівняння регресії лише змінних, що мають достовірний рівень значущості (рис. 84).

		Regression Summary for Dependent Variable: Біологічний вік (Spreadsheet1)					
		R= .70164330 R <sup>2</sup> = .49230331 Adjusted R <sup>2</sup> = .40997412					
		F(6, 37)=5,9797 p<.00019 Std. Error of estimate: 4,2822					
N=44		b*	Std. Err. of b*	b	Std. Err. of b	t(37)	p-value
Intercept				60,10376	11,14138	5,39464	0,000004
Індекс СО <sub>2</sub> , бал		0,348338	0,126385	0,56821	0,20616	2,75617	0,009024
Індекс маси тіла, кг/м <sup>2</sup>		-0,366825	0,134788	-1,09692	0,40306	-2,72149	0,009849
Життєвий індекс, мл/кг		-0,499296	0,133370	-0,36835	0,09839	-3,74369	0,000615
Індекс Робінсона		0,421491	0,136236	0,14054	0,04543	3,09384	0,003750
Силовий індекс, %		-0,051360	0,123708	-0,04096	0,09866	-0,41517	0,680417
Час відновлення, с		-0,172796	0,130157	-0,03013	0,02269	-1,32760	0,192445

Рис. 84. Вікно результатів регресійного аналізу

Натискаємо кнопку **Cancel (Закрити)** та повертаємось до меню вибору змінних (рис. 85).

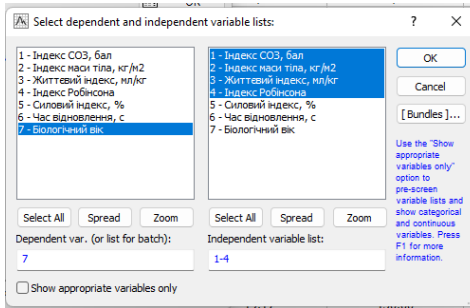


Рис. 85. Вікно вибору змінних

Натиснути кнопку **OK** для продовження, аналізуємо отримані результати (рис. 86).

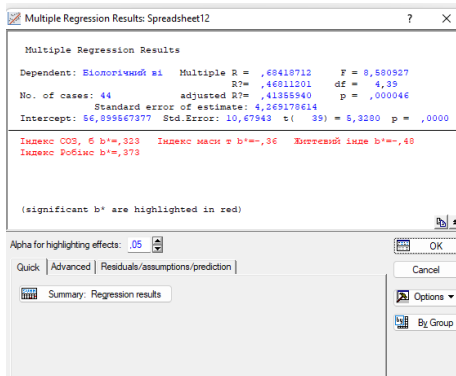


Рис. 86. Діалогове вікно регресійного аналізу

Натиснути кнопку **Summary: Regression results**. 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. 87).

Regression Summary for Dependent Variable: Біологічний вік (Spreadsheet12)						
R= .68418712 R²= .46811201 Adjusted R²= .41355940						
F(4,39)=8,5809 p<.00005 Std.Error of estimate: 4,2692						
N=44	b*	Std Err. of b*	b	Std Err. of b	t(39)	p-value
Intercept			56,89957	10,67943	5,32796	0,000004
Індекс СО2, бал	0,323275	0,124296	0,52733	0,20275	2,60084	0,013071
Індекс маси тіла, кг/м²	-0,363907	0,134000	-1,08820	0,40070	-2,71572	0,009804
Життєвий індекс, мл/кг	-0,481537	0,132294	-0,35525	0,09760	-3,63990	0,000790
Індекс Робінсона	0,372771	0,130635	0,12430	0,04356	2,85353	0,006886

Рис. 87. Вікно результатів регресійного аналізу

Коефіцієнт детермінації дорівнює 0,47, що свідчить про відсоток поясненої дисперсії залежної змінної незалежними (факторними) змінними – 47%.

Перевірка коефіцієнта детермінації за критерієм Фішера у ході дисперсійного аналізу показала, що він має високий рівень значущості  $p < 0,001$ . Цей факт підтверджує адекватність даної регресійної моделі (рис. 88).

Summary Statistics; DV: Біологічний вік (Spreadsheet12)	
Statistic	Value
Multiple R	0,684187116
Multiple R²	0,468112009
Adjusted R²	0,413559395
F(4,39)	8,58092714
p	0,0000456063462
Std.Err. of Estimate	4,26917861

Рис. 88. Вікно результатів регресійного аналізу (підсумкова статистика)

Переходимо до складання рівняння регресії. У нижній таблиці результатів у стовпці **Коефіцієнти** вказані значення  $b_0$  та параметри рівняння регресії для незалежних змінних. Значення  $b_0$ , знаходиться як **Y-перетин**, нижче у таблиці – значення незалежних змінних  $X_1 - X_4$ . Для даних значень рівняння регресії має вигляд:

$$\hat{y} = 56,90 + 0,53x_1 - 1,09x_2 - 0,36x_3 + 0,12x_4$$

Дане рівняння являє собою математичну модель регресійної залежності між показниками біологічного віку та індексу СО2 ( $x_1$ ), індексу маси тіла ( $x_2$ ), життєвого індексу ( $x_3$ ) та індекс Робінсона ( $x_4$ ).

### Практична частина.

Задання для практичної роботи представлені в Практичній роботі №10, Практична частина.



## Практична робота № 12

### Тема «Факторний аналіз у програмі Statistica»

**Мета заняття:** ознайомитися із методом факторного аналізу та способами розрахунку в програмі Statistica

#### Зміст і хід заняття

1. Факторний аналіз у програмі Statistica.

Виконання та захист практичного завдання.

#### 1. Факторний аналіз у програмі Statistica

Завданням факторного аналізу є об'єднання великої кількості показників, ознак, якими характеризується певний процес або об'єкт, у меншу кількість штучно побудованих на їх основі факторів, щоб отримана в результаті система факторів (що так само добре описує вибірккові дані, що і вихідна) була найбільш зручна з погляду змістовної інтерпретації. Чинники ви не спостерігаєте, вони присутні гіпотетично!

Statistica. Перевірка результатів контрольного прикладу у Factor Analysis

Є система змінних  $Y_1 \dots Y_3, X_4 \dots X_{17}$ .

Значення змінних чи ознак кожного зі спостережень відомі. Наявна інформація представлена у вигляді матриці, кожен рядок якої складається зі значень одного показника для кожного з об'єктів дослідження. Передбачається, що кожен елемент цієї матриці є результатом впливу деякого числа загальних гіпотетичних факторів і одного характерного фактору, тобто являють собою лінійну комбінацію неспостережуваних, гіпотетичних, безпосередньо не вимірюваних факторів. Отже, будь-який метод факторного аналізу має одне головне завдання: представити результуючий фактор у вигляді лінійної комбінації деякої кількості загальних факторів та одного характерного фактору.

Розглянемо процедуру вирішення практичного завдання методом факторного аналізу у системі STATISTICA.

**Приклад 1.** В групі юних спортсменів проведено комплексне тестування фізичних здібностей. На основі проведеного тестування необхідно визначити факторну структуру фізичної підготовленості та відібрати найбільш інформативні тести. Таблиця з вихідними даними представлена на рисунку 89.

Створіть новий файл – для цього в меню **Файл** вибрати пункт **New**.

Створіть таблицю з вихідними даними в Statistica.

	1	2	3	4	5	6	7	8	9
	підтягуван	стрибок у	стрибок уг	вис на зігн	біг 20 м, с	підніманн	проба Ром	стрибок з г	нахил тулу
1	15	205	33	28	3,5	37	23	690	20
2	16	225	43	28	4	47	12	540	15
3	20	206	31	42	4	41	10	420	22
4	21	208	32	53	3,5	48	18	460	20
5	6	151	29	16	3,9	40	10	330	5
6	24	200	30	50	3,6	49	26	360	23
7	18	200	31	38	3,8	44	25	410	20
8	25	180	29	29	3,7	49	11	360	28
9	20	224	34	56	3,6	46	24	420	21

Рис. 89. Таблиця з вихідними даними

У меню **Statistics (Статистики)** вибрати пункт **Mult/Exploratory (Множинні дослідницькі методи)**, далі обрати пункт меню **Factor (Факторний)** (рис. 90).

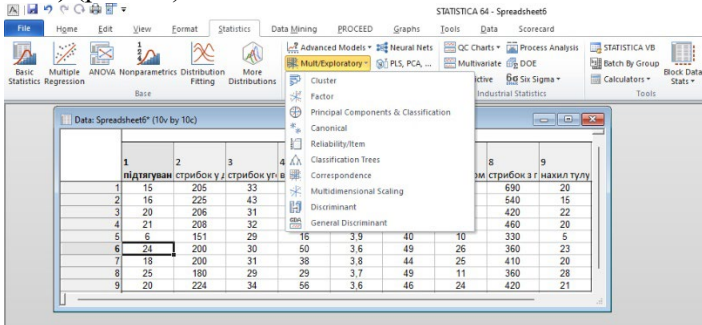


Рис. 90. Вибір факторного аналізу з горизонтального меню

У новому вікні натиснути кнопку **Variabeles (Змінні)** та обрати **Select All (Обрати усі)**, після чого натиснути **OK** (рис. 91).

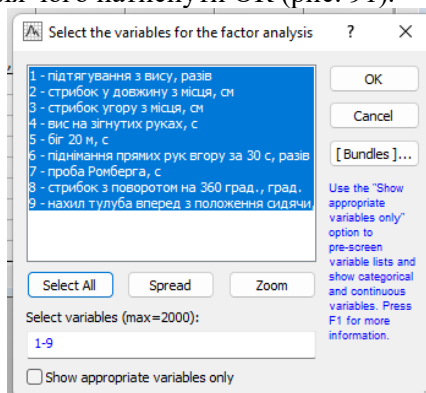


Рис. 91. Вибір змінних

У рядку **Input File (Файл вхідних даних)** вкажіть тип вихідного файлу, з яким ви будете працювати.

У модулі можливі такі типи вихідних даних:

**Correlation Matrix (Кореляційна матриця);**

**Raw Data (Вихідні дані).**

Виберіть, наприклад, Raw Data. Це звичайний файл даних, де рядками записані значення змінних.

У рядку **Missing Data (Пропущені дані)** вкажіть спосіб обробки пропущених значень.

**Casewise** – спосіб виключення пропущених випадків (полягає в тому, що в електронній таблиці, що містить дані, ігноруються всі рядки (випадки), в яких є хоча б одне пропущене значення. Це стосується всіх змінних. У таблиці залишаються лише випадки, в яких немає жодного пропуску).

**Pairwise** – парний спосіб виключення пропущених значень (ігноруються пропущені випадки не для всіх змінних, а лише для обраної пари. Випадки, в яких немає пропусків, використовуються в обробці, наприклад, при поелементному обчисленні кореляційної матриці, коли послідовно розглядаються всі пари змінних). Очевидно, у способі **Pairwise** залишається більше спостережень для обробки, ніж у способі **Casewise**. Проте, особливість у тому, що у **Pairwise** оцінки різних коефіцієнтів кореляції будуються за різним числом спостережень.

**Mean Substitution** – підстановка середнього замість пропущених значень.

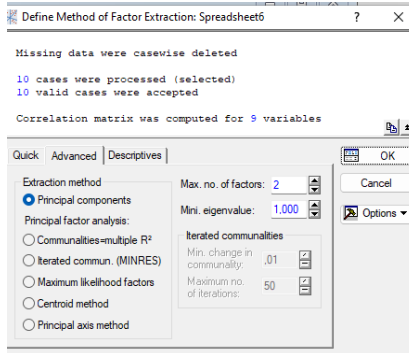
Виберіть **Casewise** – спосіб виключення пропущених випадків.

Клацнувши у стартовому вікні модуля на кнопку ОК, ви розпочнете аналіз вибраних змінних.

STATISTICA обробить пропущені значення тим способом, який ви вказали, обчислить кореляційну матрицю і запропонує на вибір кілька методів факторного аналізу.

Обчислення кореляційної матриці (якщо вона не задається відразу) – перший етап факторного аналізу.

Натисніть кнопку ОК, перед вами з'явиться вікно **Define Method of Factor Extraction (Визначити метод виділення факторів)**. Дане вікно має таку структуру (рис. 92).



*Рис. 92. Вікно Define Method of Factor Extraction (Визначити метод виділення факторів)*

Перша (верхня) частина вікна є інформаційною: тут повідомляється, що пропущені значення оброблені методом **Casewise**. Оброблено 10 випадків, та 10 випадків прийняті для подальших обчислень. Кореляційна матриця обчислена для перших 9.

Друга (нижня) частина діалогового вікна **Define Method of Factor Extraction (Визначити метод виділення факторів)**, містить 4 функціональні кнопки та опції вибору методу – ліва частина вікна, а також поля, в яких проводяться установки для ітеративного обчислення спільностей.

Зверніть увагу на поля у правій частині вікна:

**Maximum no.of factors** – Максимальна кількість факторів;

**Minimum eigenvalue** – Мінімальне значення.

Ці поля визначають кількість чинників, виділених системою. Власні значення менше зазначеного в полі ігноруються.

Розглянемо функціональні кнопки вікна. З 4 функціональних кнопок 2 – стандартні: кнопка **OK** натискається для запуску обчислювальної процедури, коли вибрано метод та проведено установки на полях. Кнопка **Cancel (Скасувати)** припиняє роботу у вікні та повертає у стартове вікно модуля.

Ініціювавши кнопку **Review correlations/means/standart diviation (Переглянути кореляції/середні/стандартні відхилення)**, ви відкриєте вікно **Переглянути описові статистики**.

За допомогою кнопки **Compute multiple regression analyses (Виконати множинну регресію)** ви можете виконати множинну регресію, не виходячи з модуля.

Група опцій, що об'єднані під заголовком **Extraction method** – метод виділення, дозволяє вибрати метод обробки.

Як мовилося в математичному аналізі, залежно від критерію оптимальності можливий аналіз або методом **Principal components – методом головних компонент**, або одним із способів, об'єднаних у групу **Principal factor analysis – аналіз основних (загальних) чинників**.

Система пропонує такі методи групи **Principal factor analysis – аналіз основних факторів**:

Communalities = multiple  $R^2$  (спільності дорівнюють квадрату коефіцієнту множинної кореляції);

Iterated communalities (MINRES) – метод ітеративних спільностей (мінімальних залишків);

Centroid method – центроїдний метод;

Principal axis method – спосіб основних осей.

У вікні **Define Method of Factor Extraction (Визначити метод виділення факторів)** натисніть кнопку **Review correlations/means/standart diviation (Переглянути кореляції/середні/стандартні відхилення)**.

Перед вами з'явилося вікно перегляду описових статистик для даних, де можна подивитися середні, стандартні відхилення, кореляції, коваріації, побудувати різні графіки (рис. 93).

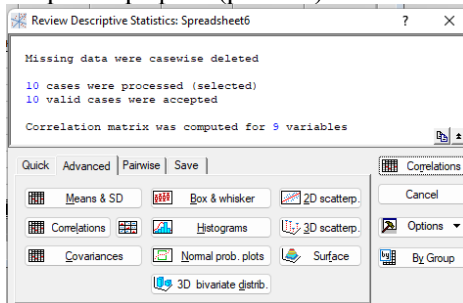


Рис. 93. Вікно перегляду описових статистик

Натисніть кнопку **Correlations (Кореляції)**. Ви побачите на екрані кореляційну матрицю обраних раніше змінних (рис. 94).

Variable	підтягування з вису, разів	стрибок у довжину з місця, см	стрибок угору з місця, см	вис на зігнутих руках, с	біг 20 м, с	підняття прямих рук угору за 30 с, разів	проба Ромберга, с	стрибок з поворотом на 360 град., град.	нахил тулуба вперед з положення сидіння, см
підтягування з вису, разів	1,00	0,48	-0,03	0,69	-0,60	0,73	-0,01	0,02	-0,96
стрибок у довжину з місця, см	0,48	1,00	0,68	0,64	-0,25	0,30	0,27	0,53	0,45
стрибок угору з місця, см	-0,03	0,68	1,00	-0,01	0,13	0,15	-0,11	0,55	-0,07
вис на зігнутих руках, с	0,69	0,64	-0,01	1,00	-0,52	0,53	0,37	-0,04	0,58
біг 20 м, с	-0,60	-0,25	0,13	-0,52	1,00	-0,33	-0,16	-0,35	-0,65
підняття прямих рук угору за 30 с, разів	0,73	0,30	0,15	0,53	-0,33	1,00	-0,06	-0,33	0,54
проба Ромберга, с	-0,01	0,27	-0,11	0,37	-0,16	-0,06	1,00	0,08	-0,03
стрибок з поворотом на 360 град., град.	0,02	0,53	0,55	-0,04	-0,35	-0,33	0,08	1,00	0,18
нахил тулуба вперед з положення сидіння, см	0,96	0,45	-0,07	0,58	-0,65	0,54	-0,03	0,18	1,00

Рис. 94. Кореляційна матриця

Збережіть кореляційну матрицю, ініціювавши кнопку **Save correlations (Зберегти кореляції)**. Далі в модулі Факторний аналіз при подальших дослідженнях можна одразу працювати з кореляційною матрицею.

Клацніть кнопку **Cancel (Закрити)** та поверніться до вікна **Define Method of factor Extraction (Визначити метод виділення факторів)**.

Виберіть опцію **Principal components (Головні компоненти)** та натисніть кнопку **OK**.

Система швидко здійснить обчислення, і на екрані з'явиться вікно **Factor Analysis Results (Результати факторного аналізу)** (рис. 95).

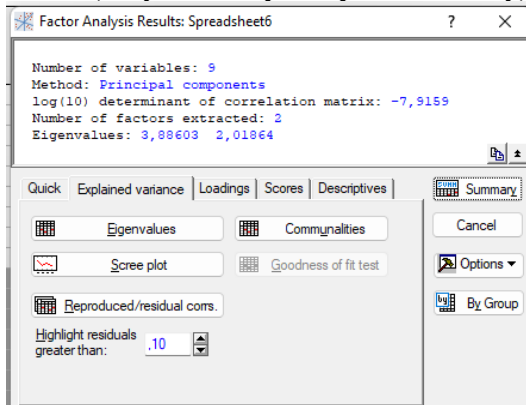


Рис. 95. Вікно *Factor Analysis Results (Результати факторного аналізу)*

У верхній частині вікна **Результати факторного аналізу** надається інформаційне повідомлення:

**Number of variables** (число аналізованих змінних) – 9;

**Method** (метод аналізу – основні компоненти);

**log(10) determination of correlation matrix** (десятковий логарифм детермінанта кореляційної матриці) – 7,9189;

**Number of Factor extraction** (кількість виділених факторів) – 2;

**Eigenvalues** (власні значення) – 3,88603; 2,01864.

У нижній частині вікна знаходяться функціональні кнопки, що дозволяють всебічно переглянути результати аналізу чисельно та графічно.

Виникає питання: скільки факторів слід обирати для певної сукупності змінних? Для цього в програмному пакеті Statistica існує критерій **Scree plot** (Критерій кам'янистого осипу).

У вікні **Factor Analysis Results** натисніть кнопку **Scree plot** отримайте наступний графік (рис. 96).

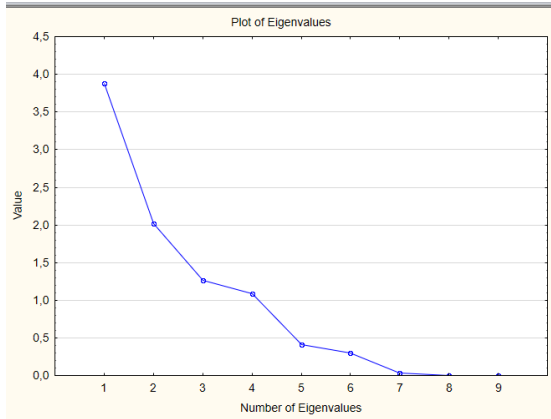


Рис. 96. Вікно Scree plot

Для визначення кількості факторів, які доцільно обирати для побудови факторної моделі, використовують один із критеріїв:

- **Критерій Кайзера** – згідно з яким число факторів дорівнює числу компонентів, власні значення яких більші за 1;
- **Критерій відсіювання Кеттелла** – згідно з яким кількість факторів визначається приблизно за точкою вигину на графіку власних значень до його виходу на пологу пряму.

Для визначення кількості факторів скористаємось критерієм Кайзера. Оскільки кількість компонентів із власним значенням більше 1 дорівнює 4, то обираємо 4 фактори для побудови структури.

Натисніть кнопку **Cancel (Закрити)** та поверніться до вікна **Define Method of factor Extraction (Визначити метод виділення факторів)**. У пункті **Max. no. of factors (Максимальна кількість факторів)** оберіть 4. Натисніть **OK** (рис. 97).

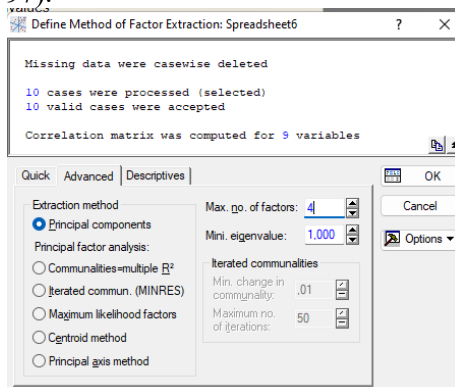


Рис. 97. Вибір кількості факторів

Оберіть вкладку **Loadings (Навантаження)** та в полі **Factor rotation (Обертання факторів)**. Ініціюйте кнопку **Varimax raw (Варімакс вихідний)** (рис. 98).

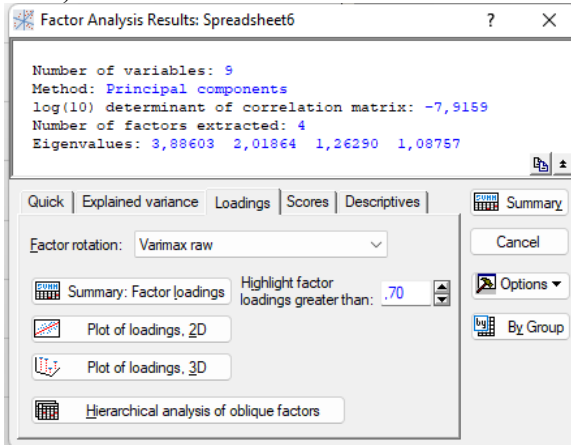


Рис. 98. Вибір функції обертання факторів

У цьому вікні можна вибрати різні повороти осі. Вікно пропонує кілька можливостей оцінити та знайти потрібний поворот такими методами:

**Varimax** – Варімакс;

**Biquartimax** – Біквартімакс;

**Quartimax** – Квартімакс;

**Equamax** – Еквімакс.

Додатковий термін у назві методів – **normalized (нормалізовані)** – вказує на те, що факторні навантаження у процедурі нормалізуються, тобто діляться на квадратний корінь з відповідної дисперсії. Термін **raw (вихідні)** показує, що навантаження, що обертаються, не нормалізовані.

Ініціюйте кнопку **Varimax raw (Варімакс вихідний)**.

Система здійснить обертання факторів методом варімаксу, і вікно **Factor Analysis Results (Результати факторного аналізу)** знову з'явиться на моніторі.

Мета обертання – отримання простої структури, коли більшість спостережень знаходиться поблизу осей координат. За випадкової конфігурації спостережень неможливо отримати просту структуру.

Клацніть на кнопці **Factor Loadings (Факторні навантаження)** і перегляньте навантаження чисельно (рис. 99).



Factor Loadings (Varimax raw) (Spreadsheet6)				
Extraction: Principal components				
(Marked loadings are >.700000)				
Variable	Factor 1	Factor 2	Factor 3	Factor 4
підтягування з вису, разів	0,967177	0,109398	0,004633	-0,108669
стрибок у довжину з місця, см	0,405430	0,813537	0,336068	0,156018
стрибок угору з місця, см	-0,098113	0,956718	-0,142897	0,139886
вис на зігнутих руках, с	0,697544	0,205287	0,542105	-0,197140
біг 20 м, с	-0,767141	0,162770	-0,162961	-0,434114
підняття прямих рук угору за 30 с, разів	0,690833	0,243291	-0,062710	-0,571744
проба Ромберга, с	-0,033919	-0,013444	0,955670	0,066803
стрибок з поворотом на 360 град., град.	0,062825	0,457772	0,018825	0,877923
нахил тулуба вперед з положення сидячи, см	0,942838	0,036555	-0,035722	0,116180
Expl. Var	3,555796	1,928020	1,372686	1,398638
Prp.Totl	0,395088	0,214224	0,152521	0,155404

Рис. 99. Факторні навантаження

Зверніть увагу на навантаження параметрів більше 0,70 (в програмі виділені червоним кольором). Наразі знайдене рішення вже можна інтерпретувати.

Фактори частіше інтерпретують за навантаженнями. Перший фактор найтісніше пов'язаний з  $X_1$  – підтягування в висі (силові якості),  $X_9$  – нахил тулуба вперед із положення сидячи (гнучкість хребта),  $X_5$  – біг на 20 м (швидкісні якості). Другий фактор найтісніше пов'язаний з  $X_3$  – стрибок угору з місця (швидкісно-силові якості),  $X_2$  – стрибок у довжину з місця (швидкісно-силові якості). Третій фактор пов'язаний тільки з одним показником –  $X_7$  – проба Ромберга (статична рівновага). Четвертий фактор також пов'язаний тільки з одним показником –  $X_8$  – стрибок з поворотом на 360 град., град. (координація).

Натисніть кнопку **Eigenvalues (Власні значення)** для отримання інформації щодо відсотку дисперсії, що пояснює кожен із отриманих факторів (рис. 100).

Eigenvalues (Spreadsheet6)				
Extraction: Principal components				
Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,886033	43,17815	3,886033	43,17815
2	2,018636	22,42928	5,904669	65,60743
3	1,262898	14,03219	7,167566	79,63963
4	1,087573	12,08415	8,255140	91,72378

Рис. 100. Власні значення та відсоток поясненої дисперсії факторів

В таблиці бачимо, що перший фактор пояснює 43,18% від загальної дисперсії, другий фактор – 22,43%, третій фактор – 14,03%, четвертий фактор – 12,08%. Загальний відсоток поясненої дисперсії в отриманій факторній структурі – 91,72%.

### Практична частина.

1. У групі жінок першого зрілого віку проведено тестування показників фізичного розвитку та рівня розвитку гнучкості. Педагогічний експеримент був спрямований на визначення ефективності комплексної програми оздоровчого стрейчинг-тренування. Визначити факторну структуру за наведеними статистичними ознаками на початку та в кінці дослідження, проаналізувати отримані результати.

Дані тестування на початку педагогічного експерименту

Маса тіла, кг	ЖЄЛ, л	Життєвий індекс, ум.од.	Згинання плеча, град.	Відведення плеча, град.	Згинання стегна, град.	Відведення стегна, град.	Згинання в колін. суглобі, град.
66	1,9	29	170	168	120	20	110
60	2,3	38	176	175	130	30	119
64	1,5	28	160	167	110	35	100
63	2,5	40	165	165	89	20	93
51	2,2	43	166	178	120	25	113
57	2,2	38	172	175	90	20	100
60	2,0	33	180	164	100	60	134
64	2,6	41	178	178	128	50	101
81	2,4	37	167	180	138	60	150
51	2,8	55	178	178	120	50	110
65	2,5	38	180	140	110	35	110
58	2,5	43	143	148	110	32	140
63	2,3	36	158	136	118	30	90
70	2,5	36	158	160	118	52	137
61	2,0	33	178	177	122	25	90
58	2,3	40	175	175	127	55	110
60	1,8	30	175	170	112	20	128
76	2,8	37	175	165	109	30	123
46	2,5	54	163	162	132	50	153
70	1,81	26	139	135	120	45	110

## Дані тестування в кінці педагогічного експерименту

Маса тіла, кг	ЖСЛ, л	Життєвий індекс, у.од.	Згинання плеча, град.	Відведення плеча, град.	Згинання стегна, град.	Відведення стегна, град.	Згинання в колін. суглобі, град.
62	2,1	34	170	168	120	20	110
57	2,5	44	177	177	130	32	120
61	1,9	31	160	167	110	35	100
59	2,5	42	167	168	90	20	90
51	2,2	43	166	178	120	25	113
55	2,4	43	173	177	90	21	100
58	2,4	41	180	166	110	60	136
70	2,6	37	180	178	130	50	101
75	2,8	37	167	180	140	60	150
51	2,8	55	178	178	120	50	110
67	2,7	39	180	140	110	35	110
54	2,6	48	143	148	110	32	140
59	2,5	42	158	136	120	30	90
65	2,8	43	160	160	120	52	137
60	3,0	50	177	180	123	25	90
64	2,5	39	180	178	130	55	120
54	2,3	42	177	170	114	20	130
69	3,1	45	176	167	110	30	125
48	2,7	55	168	164	132	50	150
64	2,0	31	140	135	120	43	110

2. У групі спортсменів досліджували психологічні риси особистості. Проаналізувати результати на початку та в кінці дослідження за факторною структурою даних.

## Дані тестування на початку дослідження

Мотивація на успіх	Мотивація на невдачі	Особистісна тривожність	Ситуативна тривожність	Самопочуття	Активність	Настрій
1	2	3	4	5	6	7
16	16	31	23	6,4	2,1	6,8
15	12	44	28	5,5	5,3	5,7
23	10	55	45	5,5	4,7	5,1
18	17	48	46	4,9	4,1	3,1
19	12	52	59	5,5	5,8	6,1

1	2	3	4	5	6	7
16	23	41	48	5,2	5,4	5,3
16	9	50	42	6,7	6,9	6,3
20	18	51	57	5,3	6,4	4,7
15	13	42	47	4,4	2,9	6
22	12	33	44	5,2	6,2	4,7
15	21	45	41	5,6	5,3	5,6
20	22	43	41	5,7	6,7	4,4
21	23	44	47	5,3	5,1	5,1
21	11	30	36	6,7	4	7
23	10	46	64	5,7	5,7	3,7
17	21	54	54	6,2	6	5,8
23	17	54	42	6,1	2,1	6,1
22	14	41	35	6	6,4	5,7
21	10	27	34	6,1	6,4	6,1

Дані тестування в кінці дослідження

Мотивація на успіх	Мотивація на невдачі	Особистісна тривожність	Ситуативна тривожність	Самопоцугтя	Активність	Настрій
20	14	30	33	6,9	4,4	6,7
16	13	37	31	6,4	6,1	6,8
21	13	55	43	4,8	5,2	4,2
22	14	58	46	4,9	4,1	3,1
18	15	50	54	4,7	4,9	6,2
22	18	41	46	5,1	5,9	6,6
22	14	41	42	6,8	6,7	7
17	13	41	35	4,9	4,8	6,3
20	16	45	48	5,5	4,9	6,8
21	5	39	32	5,9	5	6,1
16	20	48	56	5,3	5,2	5,1
19	17	48	48	6	5,5	6
16	21	42	41	3,4	4,6	4,5
19	12	33	35	5,8	5,6	6,8
19	9	46	56	5,9	5,9	6
23	17	49	54	6,1	5,5	5,6
22	13	48	42	5,7	5,7	5,7
22	17	49	44	5,8	5,2	5,8
17	13	34	35	5,9	5,1	6,4

3. Визначити та проаналізувати факторну структуру показників фізичного розвитку та фізичної підготовленості групи дітей.

Довжина тіла, см.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Маса тіла, кг.	128	25	94	112	65	2	35	29	105,3	14,9	53,2	10	12	121	25	26	32	6
ЧСС сп.	121	20	90	100	60	1,6	30	22	90	14,7	59,8	11	9	115	19	40	34	7
АТ сист.	125	24	96	110	64	1,5	34	30	105,6	15,2	63,4	11	10	98	13	58	31	4
АТ діаст.	125	22	97	112	60	1,6	37	34	108,6	16,2	44,1	8	9	123	21	43	19	4
ЖСЛ	127	24	98	118	65	1,4	30	28	115,6	15,5	52,9	10	10	110	11	36	27	5
Проба Штанге	125	21	96	115	63	1,6	32	30	110,4	15,8	54,6	9	8	118	22	23	23	6
Проба Гечя	123	22	95	108	62	1,7	35	31	102,6	15,3	52,5	11	12	104	13	50	34	4
Індекс Робінсона	124	22	90	110	65	1,6	33	27	99	16,2	45,6	7	6	101	15	38	26	3
Човниковий біг, с	120	21	92	106	64	1,3	30	23	97,5	15,8	54,2	8	7	97	17	20	38	4
Паралельний степ-тест	125	24	94	114	68	1,5	30	26	107,1	16,4	57,4	10	10	112	18	44	43	5
Динамометрія пр.	124	22	93	110	65	1,6	32	27	104,9	15,7	56	10	11	106	15	25	22	6
Динамометрія лів.	122	23	100	116	68	1,6	32	25	102,3	16,3	54,7	9	9	103	14	29	19	4
Вис, с	124	22	94	110	64	1,2	28	19	116	15,3	55,8	10	9	105	20	40	23	4
Підлімання тулуба, разів	119	22	89	114	60	1,1	27	24	101,4	15,2	50,4	7	6	87	10	25	20	5
Рівновага, с	122	23	92	118	62	1,5	34	29	108,5	16,2	48	7	6	94	8	40	31	7
Пунктирність, см																		

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
121	22	91	116	62	1.2	29	21	105,5	15,5	50,2	9	9	112	11	30	23	5
124	21	94	118	64	1,4	35	26	110,9	16,1	51,1	8	8	106	12	36	39	5
125	24	96	110	70	1,3	34	21	105,6	15,8	51,4	8	7	100	13	34	18	4
120	21	90	110	64	1,2	30	24	99	16,4	50,8	10	9	102	14	28	40	4
123	22	92	115	60	1,3	31	23	105,8	17	50	7	6	99	12	55	22	4
124	21	96	118	65	1,5	29	23	103,4	17,2	50	8	6	104	15	50	25	3
121	21	92	116	62	1,2	32	22	113,2	16,6	42,7	9	7	97	14	33	27	4
125	23	91	115	63	1,3	31	23	103,5	17,2	50,1	8	7	101	12	46	24	6
120	20	95	112	60	1	30	24	106,7	16,8	53,4	9	9	107	9	40	33	5
119	20	93	113	63	1,2	27	23	106,4	16,9	50,2	8	8	104	12	37	35	5
124	23	90	108	64	1,2	28	25	97,2	16,3	60	9	8	108	14	56	30	4
121	21	94	118	66	1,1	28	20	110,9	16,8	47,6	9	8	101	10	50	35	6
129	27	101	110	65	1,9	36	29	111,1	14,5	52,4	11	12	128	37	71	23	5
126	24	97	105	60	1,8	32	24	101,8	14,2	51,1	10	10	124	19	60	35	3
127	24	100	108	70	1	28	20	108	14,6	59,9	8	10	104	12	28	25	5
130	25	94	115	72	1,3	31	28	108,1	14,2	52,6	10	10	128	28	100	26	4
128	27	96	110	74	1,6	33	30	105,6	14,5	57,9	10	11	103	6	20	19	5
126	24	99	104	68	1,7	33	27	102,9	14,8	58,7	8	5	132	3	29	17	2
121	20	102	100	60	1,5	36	28	102	14,3	51	12	14	139	27	70	30	3
126	24	98	106	70	1,3	29	24	112	15,2	49,4	5	5	82	19	33	30	3

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
123	25	99	107	67	1	27	20	103,9	15,2	48,1	3	7	123	14	28	20	4
122	22	100	105	63	1,8	35	29	105	13,4	57,5	11	11	141	25	100	55	6
124	24	100	112	66	1,5	31	24	104,8	14,6	57,8	8	7	94	9	36	3	7
118	26	97	107	67	1,4	30	22	99	14,3	55,6	9	5	92	8	15	13	5
120	22	99	100	72	1,3	30	24	101,7	14,5	57,7	10	11	101	16	22	19	3
124	23	96	106	74	1,5	31	26	105,2	14,8	48,5	8	8	112	32	25	18	7
117	21	90	102	60	1,2	28	22	91,8	14,9	49,9	7	5	78	10	15	23	5
124	26	94	114	62	1,5	39	26	107,1	13,9	57,5	10	9	124	18	40	41	5
126	25	96	118	68	1	24	22	113,2	15,8	58,3	6	8	103	6	50	35	3
120	23	92	112	63	1,3	32	25	103	15,9	61,4	8	5	97	8	25	8	4
122	22	91	114	65	1,3	34	26	103,7	15,9	60	9	9	111	20	20	50	6
118	20	93	108	62	1,3	33	23	100,4	15	51,1	6	6	98	10	58	20	8
119	21	89	110	60	1,2	29	20	97,9	16	56,8	5	4	96	9	60	15	5
120	23	90	110	62	1,5	34	25	99	16,1	57,4	10	9	95	4	30	20	3
122	22	95	116	70	1,1	31	22	105,1	15,4	46,6	5	5	106	8	43	25	6
120	21	91	112	63	1,1	28	19	110,2	15,2	50	10	10	102	10	42	15	5
128	26	100	114	66	1	29	24	102,9	14,6	43,7	10	9	99	12	60	30	5
123	21	92	112	65	1,1	28	22	114	14,9	51,3	6	5	101	7	34	37	3
120	23	87	114	62	1,3	29	24	99,1	14,1	52,4	9	5	122	17	46	40	1
125	27	95	116	72	1,1	29	25	110,2	15,1	49,2	10	9	80	9	50	30	9

## Практичне заняття № 13

### Тема «Кластерний аналіз у програмі Statistica»

**Мета заняття:** ознайомитися із методом кластерного аналізу та способами розрахунку в програмі Statistica

#### Зміст і хід заняття

1. Практична робота в системі Statistica.  
Виконання та захист практичного завдання.

#### 1. Практична робота в системі Statistica

Розглянемо процедуру розв'язання практичного завдання методом кластерного аналізу у системі STATISTICA.

Завданням кластерного аналізу є організація спостережуваних даних у наочні структури. Для вирішення цієї задачі в кластерному аналізі використовуються такі методи: Joining (tree clustering) (ієрархічні агломеративні методи або деревоподібна кластеризація),  $K$  – means clustering (метод  $K$  середніх), Two-way joining (двовходове об'єднання).

У Statistica реалізовані класичні методи кластерного аналізу, включаючи методи  $k$ -середніх, ієрархічної кластеризації та двовходового об'єднання.

Дані можуть надходити як у вихідному вигляді, і у вигляді матриці відстаней між об'єктами.

Спостереження та змінні можна кластеризувати, використовуючи різні міри відстані (евклідова, квадрат евклідова, манхеттенівське, Чебишева та ін.) та різні правила об'єднання кластерів (одинарний, повний зв'язок, незважене та зважене попарне середнє за групами та ін.).

Постановка задачі

Вхідний файл даних містить наступну інформацію про спортсменів:

- довжина тіла – перша змінна;
- маса тіла – друга змінна;
- довжина руки – третя змінна;
- довжина ноги – четверта змінна (рис. 101).



	1	2	3	4
	Довжина тіла	Маса тіла	Довжина руки	Довжина ноги
1	179,00	76,00	81,00	94,00
2	191,00	84,00	80,70	101,00
3	185,00	82,00	84,30	100,00
4	196,00	96,00	88,00	103,00
5	183,00	78,00	80,50	99,00
6	197,00	85,00	94,00	105,00
7	187,00	72,00	82,70	97,00
8	196,00	96,00	88,00	103,00
9	183,00	82,00	84,30	100,00
10	187,00	72,00	82,70	97,00
11	182,00	92,00	81,70	96,00
12	183,00	78,00	80,50	99,00
13	191,00	84,00	80,70	101,00
14	179,00	76,00	81,00	94,00
15	182,00	92,00	81,70	96,00
16	174,00	70,00	82,30	95,00
17	187,00	72,00	82,70	97,00
18	185,00	76,00	82,00	99,00
19	189,00	72,00	79,00	96,50
20	173,00	66,00	76,00	86,00
21	179,00	76,00	81,00	94,00

Рис. 101. Вхідний файл даних

Метою даного аналізу є розбиття спортсменів на класи, кожен із яких відповідає певному соматичному типу. Використання кластерного аналізу для вирішення цієї задачі найбільш ефективно.

У загальному випадку кластерний аналіз призначений для об'єднання деяких об'єктів на класи (кластери) таким чином, щоб в один клас потрапляли максимально схожі, а об'єкти різних класів максимально відрізнялися один від одного. Кількісний показник подібності розраховується заданим способом виходячи з даних, що характеризують об'єкти.

#### Масштаб вимірів

Усі кластерні алгоритми потребують оцінок відстаней між кластерами чи об'єктами, і зрозуміло, що з обчисленні відстані необхідно задати масштаб вимірів.

Оскільки різні вимірювання використовують абсолютно різні типи шкал, дані необхідно стандартизувати. Для цього в меню **Data (Дані)** оберіть пункт **Standardize (Стандартизувати)**, так що кожна змінна матиме середнє 0 та стандартне відхилення 1 (рис. 102).

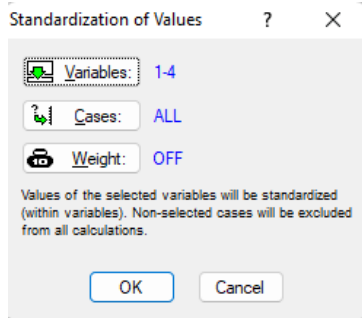


Рис. 102. Меню Стандартизація даних

Таблиця зі стандартизованими змінними наведена нижче (рис. 103).

	1	2	3	4
	Довжина тіла	Маса тіла	Довжина руки	Довжина ноги
1	-0,94	-0,45	-0,43	-0,91
2	0,87	0,48	-0,51	0,79
3	-0,04	0,25	0,45	0,55
4	1,62	1,87	1,45	1,27
5	-0,34	-0,22	-0,57	0,31
6	1,77	0,60	3,06	1,76
7	0,27	-0,91	0,02	-0,18
8	1,62	1,87	1,45	1,27
9	-0,04	0,25	0,45	0,55
10	0,27	-0,91	0,02	-0,18
11	-0,49	1,41	-0,24	-0,42
12	-0,34	-0,22	-0,57	0,31
13	0,87	0,48	-0,51	0,79
14	-0,94	-0,45	-0,43	-0,91
15	-0,49	1,41	-0,24	-0,42
16	-1,69	-1,14	-0,08	-0,66
17	0,27	-0,91	0,02	-0,18
18	-0,04	-0,45	-0,16	0,31
19	0,57	-0,91	-0,97	-0,30
20	-1,84	-1,61	-1,78	-2,84
21	-0,94	-0,45	-0,43	-0,91

Рис. 103. Таблиця зі стандартизованими змінними

На першому етапі з'ясуємо, чи утворюють спортсмени «природні» кластери, які можуть бути осмислені.

У меню **Statistics (Статистики)** виберемо пункт **Mult/Exploratory (Множинні дослідницькі методи)**, далі обрати пункт меню **Cluster (Кластерний)** для відображення стартової панелі модуля Кластерний аналіз. У цьому діалозі виберемо **Joining (tree-clustering) (Ієрархічна)** класифікація та натиснемо ОК (рис. 104).

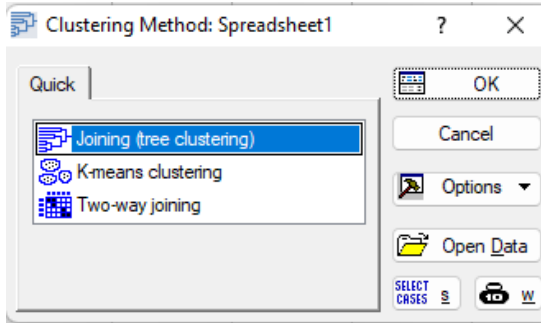


Рис. 104. Стартова панель модуля Кластерний аналіз

Натисніть кнопку **Variabeles** (Змінні), виберемо 1-4 змінні, у полі **Cluster** (Кластер) оберемо **Cases (rows)** (Спостереження) (рядки)). Правило об'єднання відзначимо Complete Linkage (Метод повного зв'язку), міру близькості – **Euclidean distances** (Евклідова відстань). Натисніть ОК (рис. 105).

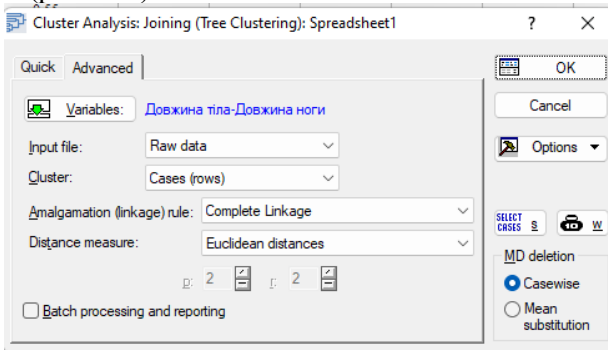


Рис. 105. Діалогове вікно налаштування параметрів кластерного аналізу

Метод повного зв'язку визначає відстань між кластерами як найбільшу відстань між будь-якими двома об'єктами в різних кластерах (тобто найбільш віддаленими сусідами).

Міра близькості, яка визначається евклідовою відстанню, є геометричною відстанню в  $n$ -мірному просторі і обчислюється наступним чином:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Виведення результатів та їх аналіз (рис. 106).

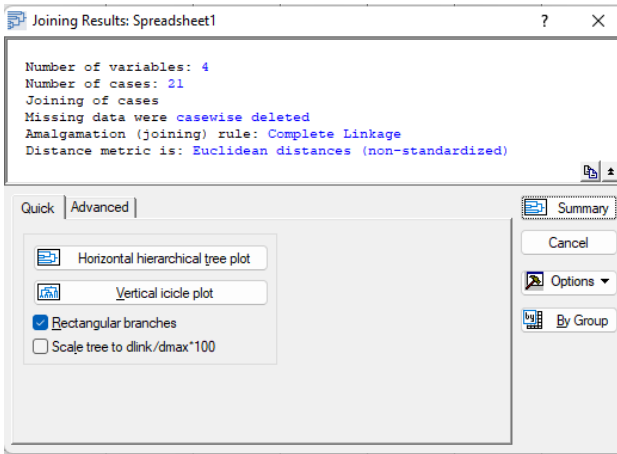


Рис. 106. Діалогове вікно *Joining Results* (Результати аналізу кластерних функцій)

Інформаційна частина діалогового вікна **Joining Results (Результати аналізу кластерних функцій)** повідомляє, що

- Number of variables – кількість змінних;
- Number of cases – кількість спостережень;
- Missing data were casewise deleted – здійснено класифікацію спостережень або змінних (залежить від рівня параметра у рядку Cluster у попередньому вікні налаштування.)
- Amalgation (joining) rule – правило об'єднання кластерів (назва ієрархічного агломеративного методу, заданого в рядку Amalgation rules, а в попередньому вікні настроювання);
- Distance.metric is – Метрика відстані (залежить від установки у рядку Distance measure у попередньому вікні налаштування).

Користувач може викликати на екран горизонтальну та вертикальну діаграму (Horizontal hierachical plot або Vertical icicle plot). Найбільш традиційне – вертикальне представлення.

Найбільш важливим результатом, що отримується в результаті деревоподібної кластеризації, є ієрархічне дерево. Натисніть кнопку Vertical icicle plot (Вертикальний графік) (рис. 107).

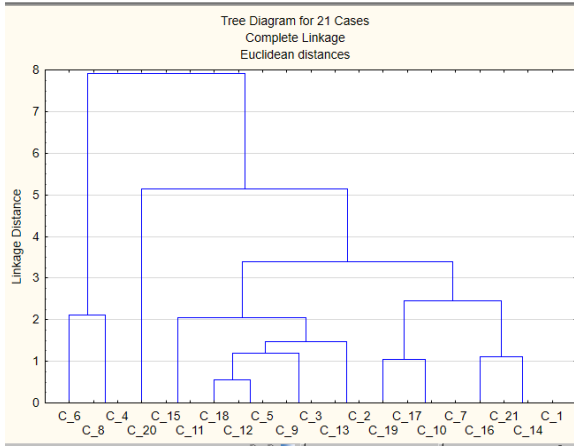


Рис. 107. Вертикальна діаграма

Спочатку деревоподібні діаграми можуть здатися трохи заплутаними, проте після деякого вивчення вони стають зрозумілішими. Діаграма починається зверху (для вертикальної дендрограми) з кожного спортсмена у своєму кластері.

Як тільки ви почнете рухатися вниз, об'єкти (спортсмени), які «тісніше стикаються один з одним», об'єднуються і формують кластери. Кожен вузол діаграми, наведеної вище, представляє об'єднання двох або більше кластерів, положення вузлів вертикальної осі визначає відстань, на якому були об'єднані відповідні кластери.

Виходячи з візуального представлення результатів, можна зробити припущення, що спортсмени утворюють три природні кластери. Перевіримо це припущення, розбивши вихідні дані методом К-середніх на 3 кластери, і перевіримо значущість різниці між отриманими групами. Для цього необхідно повернутись до Стартової панелі модуля Кластерний аналіз, послідовно натиснувши кнопку **Cancel (Закрити)** у двох попередніх діалогових вікнах.

У Стартовій панелі модуля Кластерний аналіз виберемо **K-means clustering (Кластеризація методом К-Середніх)** (рис. 108).

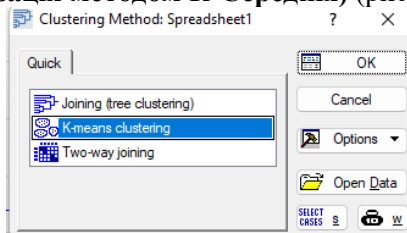
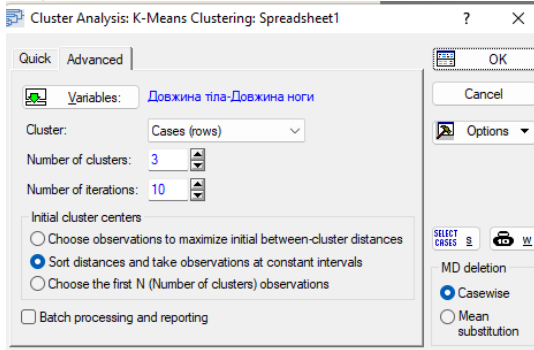


Рис. 108. Стартова панель модуля Кластерний аналіз

Натисніть кнопку **Variabes (Змінні)**, виберемо 1-4 змінні, у полі **Cluster (Кластер)** оберіть **Cases (rows) (Спостереження (рядки))**, у полі **Number of clusters (Кількість кластерів)** поставимо 3 кластери розбиття (рис. 109).



*Рис. 109. Діалогове вікно налаштування параметрів кластерного аналізу методом К-середніх*

Метод К-середніх полягає в наступному: обчислення починаються з  $k$  випадково вибраних спостережень (у нашому випадку  $k=3$ ), які стають центрами груп, після чого об'єктний склад кластерів змінюється з метою мінімізації мінливості всередині кластерів та максимізації мінливості між кластерами.

Кожне наступне спостереження ( $K+1$ ) відноситься до тієї групи, міра подібності до центру ваги якого мінімальна.

Після зміни складу кластера обчислюється новий центр тяжкості, найчастіше як вектор середніх за кожним параметром. Алгоритм триває доти, доки склад кластерів не перестане змінюватися.

Коли результати класифікації отримані, можна розрахувати середнє значення показників кожного кластера, щоб оцінити, наскільки вони різняться між собою.

У вікні **k-Means Clustering Results (Результати методу К середніх)** оберемо **Analysis of Variance (Дисперсійний аналіз)** для визначення значущості відмінності між отриманими кластерами (рис. 110).

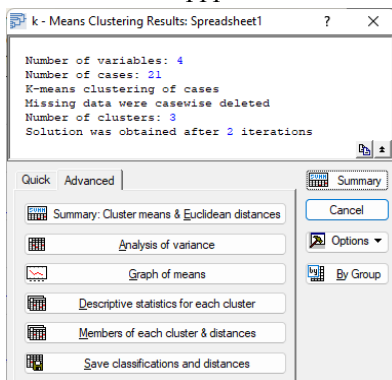


Рис. 110. Вікно *k*-Means Clustering Results (Результати методу *K* середніх)

Отже, значення  $p < 0,05$ , що говорить про значну відмінність (рис. 111).

Variable	Analysis of Variance (Spreadsheet1)					
	Between SS	df	Within SS	df	F	signif. p
Довжина тіла	16,55153	2	3,44848	18	43,19699	0,000000
Маса тіла	9,62447	2	10,37553	18	8,34851	0,002722
Довжина руки	14,45863	2	5,54137	18	23,48293	0,000010
Довжина ноги	14,21618	2	5,78382	18	22,12133	0,000014

Рис. 111. Результати дисперсійного аналізу між кластерами за усіма змінними

Натисніть кнопку **Members of each clusters & distances (Елементи кластерів та відстані)** для перегляду спостережень, що входять до кожного кластера (рис. 112-114). Опція також дозволяє відобразити евклідові відстані об'єктів від центрів (середніх значень) відповідних кластерів. Перший кластер:

Members of Cluster Number 1 (Spreadsheet1) and Distances from Respective Cluster Center Cluster contains 13 cases	
Case No.	Distance
C_2	0,577641
C_3	0,418573
C_5	0,308488
C_7	0,496289
C_9	0,418573
C_10	0,496289
C_11	0,822704
C_12	0,308488
C_13	0,577641
C_15	0,822704
C_17	0,496289
C_18	0,239905
C_19	0,667024

Рис. 112. Спостереження, що входять до першого кластеру

Другий кластер:

Members of Cluster Number 2 (Spreadsheet1) and Distances from Respective Cluster Center Cluster contains 5 cases	
Case No.	Distance
C_1	0,316919
C_14	0,316919
C_16	0,480078
C_20	1,097435
C_21	0,316919

Рис. 113. Спостереження, що входять до другого кластеру

Третій кластер:

Members of Cluster Number 3 (Spreadsheet1) and Distances from Respective Cluster Center Cluster contains 3 cases	
Case No.	Distance
C_4	0,353083
C_6	0,706165
C_8	0,353083

Рис. 114. Спостереження, що входять до третього кластеру

Отже, у кожному з трьох кластерів знаходяться об'єкти зі схожими антропометричними даними.

#### Описові статистики

Знання описових статистик у кожній групі, безумовно, є важливим для будь-якого дослідника.

Відображення статистик кожного кластера з діалогу **k-Means Clustering Results (Результати методу К середніх)** наразі не становить інтересу, оскільки дані були стандартизовані.

Натисніть кнопку **Save classification and distances (Зберегти класифікацію та відстані)**. Таблиця стандартизованих даних доповнилася інформацією про кластер, до якого належить спостереження, евклідову відстань і номер спостереження (рис. 115)

	1	2	3
	CASE_NO	CLUSTER	DISTANCE
C_1	1	2	0,32
C_2	2	1	0,58
C_3	3	1	0,42
C_4	4	3	0,35
C_5	5	1	0,31
C_6	6	3	0,71
C_7	7	1	0,50
C_8	8	3	0,35
C_9	9	1	0,42
C_10	10	1	0,50
C_11	11	1	0,82
C_12	12	1	0,31
C_13	13	1	0,58
C_14	14	2	0,32
C_15	15	1	0,82
C_16	16	2	0,48
C_17	17	1	0,50
C_18	18	1	0,24
C_19	19	1	0,67
C_20	20	2	1,10
C_21	21	2	0,32

Рис. 115. Таблиця із результатами розбиття об'єктів на кластери



Скопіюємо змінну **CLUSTER (КЛАСТЕР)** у вихідну таблицю даних (рис. 116).

	1	2	3	4	5
	Довжина тіла	Маса тіла	Довжина руки	Довжина ноги	CLUSTER
1	179,00	76,00	81,00	94,00	2
2	191,00	84,00	80,70	101,00	1
3	185,00	82,00	84,30	100,00	1
4	196,00	96,00	88,00	103,00	3
5	183,00	78,00	80,50	99,00	1
6	197,00	85,00	94,00	105,00	3
7	187,00	72,00	82,70	97,00	1
8	196,00	96,00	88,00	103,00	3
9	185,00	82,00	84,30	100,00	1
10	187,00	72,00	82,70	97,00	1
11	182,00	92,00	81,70	96,00	1
12	183,00	78,00	80,50	99,00	1
13	191,00	84,00	80,70	101,00	1
14	179,00	76,00	81,00	94,00	2
15	182,00	92,00	81,70	96,00	1
16	174,00	70,00	82,30	95,00	2
17	187,00	72,00	82,70	97,00	1
18	185,00	76,00	82,00	99,00	1
19	189,00	72,00	79,00	96,50	1
20	173,00	66,00	76,00	86,00	2
21	179,00	76,00	81,00	94,00	2

Рис. 116. Вихідна таблиця даних із зазначенням кластеру об'єктів

Тепер для кожного кластера можна визначити основні описові статистики.

У меню **Statistics (Статистика) – Basic Statistics and Tables (Основні статистики і таблиці)** виберемо опцію **Breakdown & one-way ANOVA (Угруповання та Однофакторний ДА)** (рис. 117).

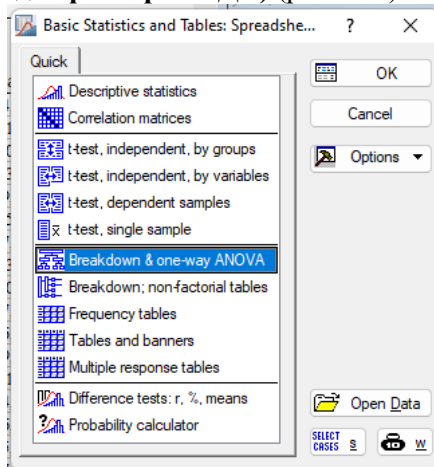


Рис. 117. Вікно *Basic Statistics and Tables (Основні статистики і таблиці)*

Обираємо залежні змінні (довжина тіла, маса тіла, довжина руки, довжина ноги) та групуючу змінну (кластер) (рис. 118).

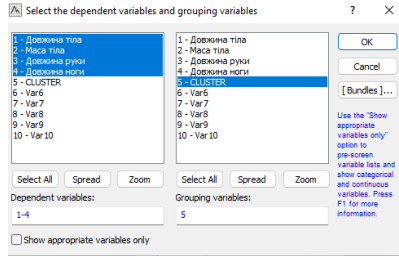


Рис. 118. Вікно вибору змінних для дисперсійного аналізу

Натискаємо кнопку **Codes of grouping variables (коди групуючої змінної)** та вибираємо **All (Усі)**. Натискаємо **OK** (рис. 119).

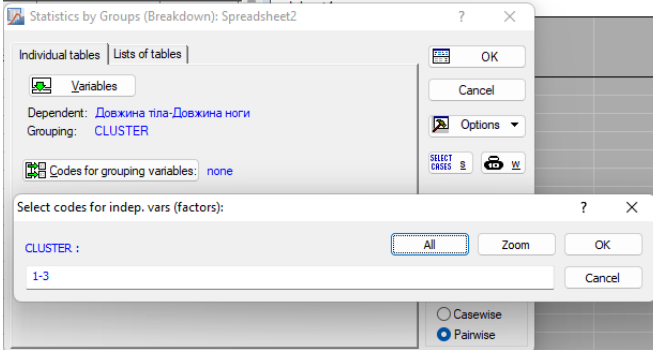


Рис. 119. Вікно кодування групуючої змінної

У вікні **Statistics by Groups – Results (Статистика для груп – Результати)** обираємо **Summary: Tables of Statistics (Підсумок: Таблиці статистики)** (рис. 120).

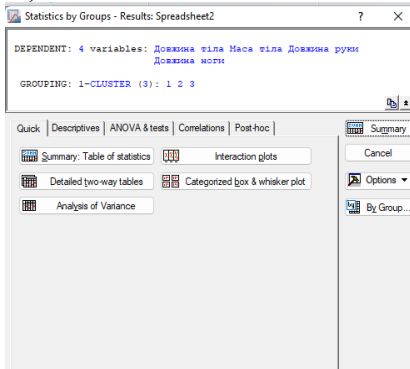


Рис. 120. Діалогове вікно дисперсійного аналізу

Нижче наведено таблиці описових статистик для кожної змінної (рис. 121):

Breakdown Table of Descriptive Statistics (Spreadsheet2) N=21 (No missing data in dep. var. list)													
CLUSTER	Довжина тіла Means	Довжина тіла N	Довжина тіла Std.Dev.	Маса тіла Means	Маса тіла N	Маса тіла Std.Dev.	Довжина руки Means	Довжина руки N	Довжина руки Std.Dev.	Довжина ноги Means	Довжина ноги N	Довжина ноги Std.Dev.	
1	185,9231	13	3,094660	79,69231	13	7,110844	81,80769	13	1,544594	98,3462	13	1,841474	
2	176,8000	5	3,033150	72,90000	5	4,604346	80,26000	5	2,447039	92,6000	5	3,714835	
3	196,3333	3	0,577350	92,33333	3	6,350853	90,00000	3	3,464102	103,6667	3	1,154701	
All Grps	185,2381	21	6,647592	79,85714	21	8,627200	82,60952	21	3,717513	97,7381	21	4,128011	

Рис. 121. Таблиця описових статистик

Побудуємо графік середніх та довірчих інтервалів для змінних у кожному кластері. Для цього активуємо кнопку **Interaction plots** (рис. 122).

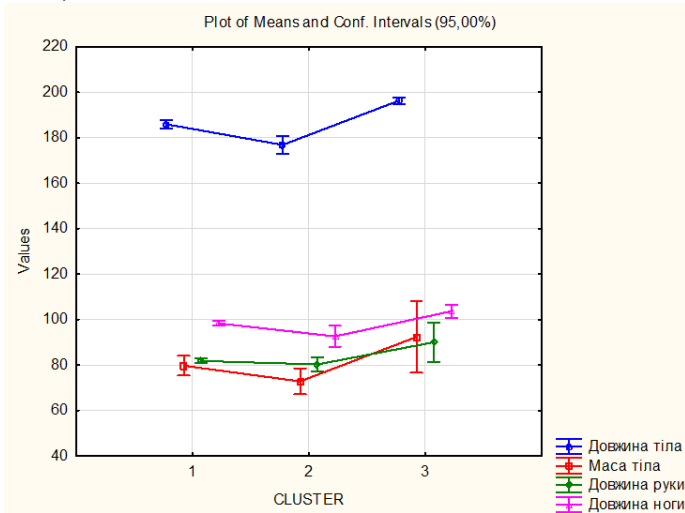


Рис. 122. Графік середніх та довірчих інтервалів для змінних у кожному кластері

Для визначення достовірності відмінностей між різними кластерами застосуємо дисперсійний аналіз, активувавши кнопку **Analysis of Variance**.

Оскільки  $p < 0,05$ , то між кластерами наявна достовірна різниця за усіма досліджуваними антропометричними показниками (рис. 123).

Analysis of Variance (Spreadsheet2)								
Marked effects are significant at $p < ,05000$								
Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Довжина тіла	731,4198	2	365,7099	152,3897	18	8,46610	43,19699	0,000000
Маса тіла	716,3355	2	358,1678	772,2359	18	42,90199	8,34851	0,002722
Довжина руки	199,8169	2	99,9084	76,5812	18	4,25451	23,48293	0,000010
Довжина ноги	242,2505	2	121,1253	98,5590	18	5,47550	22,12132	0,000014

Рис. 123. Результат дисперсійного аналізу

### Практична частина.

Проаналізувати результати спостережень за допомогою кластерного аналізу.

Проба Штанге	Проба Генча	Індекс Робінсона
35	29	105,3
30	22	90,0
34	30	105,6
37	34	108,6
30	28	115,6
32	30	110,4
35	31	102,6
33	27	99,0
30	23	97,5
30	26	107,1
32	27	104,9

### Практичне заняття № 14

#### Тема «Дискримінантний аналіз у програмі Statistica»

**Мета заняття:** ознайомитися із методом дискримінантного аналізу та способами розрахунку в програмі Statistica

#### Зміст і хід заняття

1. Дискримінантний аналіз у програмі Statistica.

Виконання та захист практичного завдання.

#### 1. Дискримінантний аналіз у програмі Statistica

##### Постановка задачі.

Метою аналізу є вивчення дискримінації між трьома спеціалізаціями в плаванні, ґрунтуючись на п'яти вимірах: довжина тіла, довжина передпліччя, висота стопи, обхват талії, обхват сідниць. Дані представлено на рисунку 124.

Загалом усі завдання дискримінантного аналізу ставлять подібне питання. Якщо ви є дослідником в галузі освіти, ви можете підставити замість "спеціалізації в спорті" "тип виключення", і замість змінних "антропометричні дані" змінні "успіхи в чотирьох ключових курсах". Якщо ви займаєтеся соціальними науками, то ви можете вивчати змінні, якими можна передбачити вибір кар'єри. При вивченні відбору персоналу вас можуть цікавити змінні, за якими можна відрізнити працівників рівня вище за середній від працівників, які не відповідають виконуваний роботі, та співробітників, які не придатні до роботи. Тому, хоча наступний приклад взято зі спорту, загальні процедури, що розглядаються тут,

застосовні значно ширше. Перший крок – це створення таблиці з вихідними даними у STATISTICA.

	1 Спеціалізація	2 Довжина тіла, см	3 Довжина передпліччя, см	4 Висота стопи, см	5 Обхват талії, см	6 Обхват сідниць, см
9	брас	179	25,5	7	86	98,5
10	брас	187	27	6,5	79,5	97
11	брас	186	27	6,5	86,5	103
12	брас	179	25,5	7	86	98,5
13	брас	187	27	6,5	79,5	97
14	брас	186	27	6,5	86,5	103
15	брас	179	25,5	7	86	98,5
16	брас	187	27	6,5	79,5	97
17	вольний стиль	182	27	6	74	95
18	вольний стиль	183	27	6,5	77	97
19	вольний стиль	181	27	8	79	95
20	вольний стиль	191	29	8,5	77	95
21	вольний стиль	182	27	6,5	69	88,5
22	вольний стиль	182	27	6	74	95
23	вольний стиль	183	27	6,5	77	97
24	вольний стиль	181	27	8	79	95
25	вольний стиль	191	29	8,5	77	95
26	вольний стиль	182	27	6,5	69	88,5
27	вольний стиль	182	27	6	74	95
28	вольний стиль	183	27	6,5	77	97
29	вольний стиль	181	27	8	79	95
30	вольний стиль	191	29	8,5	77	95
31	вольний стиль	182	27	6,5	69	88,5
32	вольний стиль	182	27	6	74	95
33	вольний стиль	183	27	6,5	77	97
34	вольний стиль	181	27	8	79	95
35	вольний стиль	191	29	8,5	77	95
36	вольний стиль	182	27	6,5	69	88,5

Рис. 124. Таблиця з вхідними даними у Statistica

У меню **Statistics (Статистики)** виберемо пункт **Mult/Exploratory (Множинні дослідницькі методи)** далі обрати пункт меню **Discriminant (Дискримінантний)** для відображення стартової панелі модуля Дискримінантний аналіз (рис. 125).

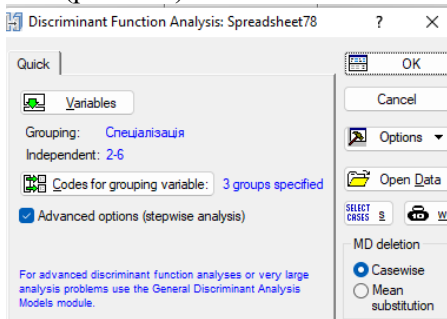


Рис. 125. Стартова панель модуля Дискримінантний аналіз

У вкладці **Quick (Швидкий)** виберіть **Advanced options (stepwise analysis) (Додаткові параметри (покроковий аналіз))**. Після натискання на кнопку **Variabeles (Змінні)** з'явиться стандартне діалогове вікно Вибір

змінних. У цьому вікні вкажіть групуючу змінну (змінна Спеціалізація) та повністю незалежні змінні, які мають бути використані для дискримінації між спеціалізаціями в плаванні.

Потім необхідно вказати коди, які були використані при групуванні змінних для ідентифікації того, до якої сукупності належить кожен спортсмен. Натисніть на кнопку **Codes for grouping variables (Коди для групуючої змінної)** та введіть 1-4, або натисніть кнопку Все, або використовуйте зірочку (\*), що відповідає відбору всіх кодів (рис. 126).

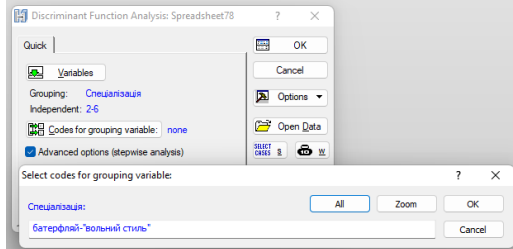


Рис. 126. Коди для групуючої змінної

Альтернативним чином, ви можете натиснути кнопку ОК на стартовій панелі, і система Statistica автоматично перегляне групуючу змінну(і), і визначить усі коди для цих змінних.

**Видалення пропущених даних.** Цей файл даних не містить жодних пропущених даних. Однак, якщо у файлі є пропущені дані, ви можете або ігнорувати спостереження з пропущеними даними (встановити в полі Видалення ПД значення Порядково) або заповнити пропущені дані середніми значеннями (Заміна середніми).

**Перегляд описових статистик.** Тепер, щоб розпочати аналіз, натисніть на кнопку ОК. Відкриється діалогове вікно **Model Definition (Визначення моделі)**, яке дозволить задати параметри дискримінантного аналізу та переглянути описові статистики (рис. 127).

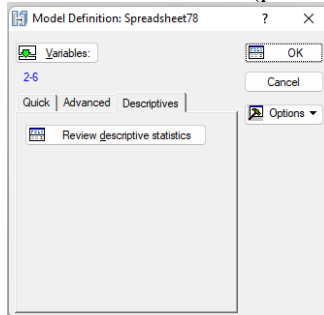


Рис. 127. Діалогове вікно Model Definition (Визначення моделі)

Перед вибором аналізу дискримінантної функції, натисніть кнопку **Review Descriptive Statistics (Переглянути описові статистики)** для того, щоб подивитися на розподіл деяких змінних та їх взаємні кореляції. Відобразиться діалогове вікно **Описові статистики** (рис. 128).

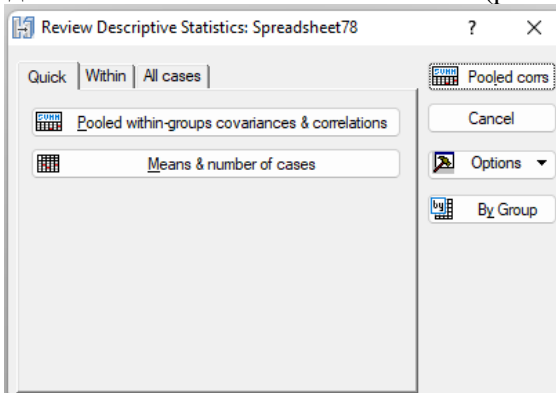


Рис. 128. Діалогове вікно Описові статистики

Спочатку погляньте на середні. У вкладці **Quick (Швидкий)** натисніть кнопку **Means & number of cases (Середні та число спостережень)**, після чого має з'явитися таблиця із середніми та діючим значенням  $n$  для кожної сукупності та для комбінації всіх груп (рис. 129).

Спеціалізація	Means (Spreadsheet78)					
	Довжина тіла, см	Довжина передпліччя, см	Висота стопи, см	Обхват талії, см	Обхват сідниць, см	Valid N
батарфляй	184.0000	26.78572	6.285714	80.00000	91.78571	7
брас	184.0000	26.50000	6.666667	84.00000	99.50000	9
вольний стиль	183.8000	27.40000	7.100000	75.20000	94.10000	20
All Grps	183.8889	27.05556	6.833333	78.33334	95.00000	36

Рис. 129. Таблиця з результатами описової статистики

**Отримання гістограми з таблиці результатів.** Для отримання гістограми частот розподілу для будь-якої змінної виділіть спочатку потрібний стовпець у таблиці. Ви можете отримати гістограму для всіх груп або тільки для вибраних груп.

Наприклад, щоб отримати гістограму змінної Довжина тіла тільки для спеціалізації Вільний стиль, пересуньте курсор на перетин другого рядка і другого стовпця таблиці. Потім натисніть праву кнопку миші, щоб відкрити контекстне меню і вибрати команду **Histogram: Entire Columns (Графіки вихідних даних)** (рис. 130).

Means (Spreadsheet78)						
Спеціалізація	Довжина тіла, см	Довжина передпліччя, см	Висота стопи, см	Обхват талії, см	Обхват сідниць, см	Valid N
батарфляй	184,0000	26,78572	6,285714	80,00000	91,78571	7
брас	184,0000	26,50000	6,666667	84,00000	99,50000	9
<b>вольний стиль</b>	<b>183,9999</b>	<b>27,40000</b>	<b>7,100000</b>	<b>75,20000</b>	<b>94,10000</b>	<b>20</b>
All Grps		27,05556	6,833333	78,33334	95,00000	36

Рис. 130. Вибір гистограми для візуалізації описової статистики

Тепер виберіть команду Гістограма – Нормальне припасування, і ви отримаєте наступний графік (рис. 131).

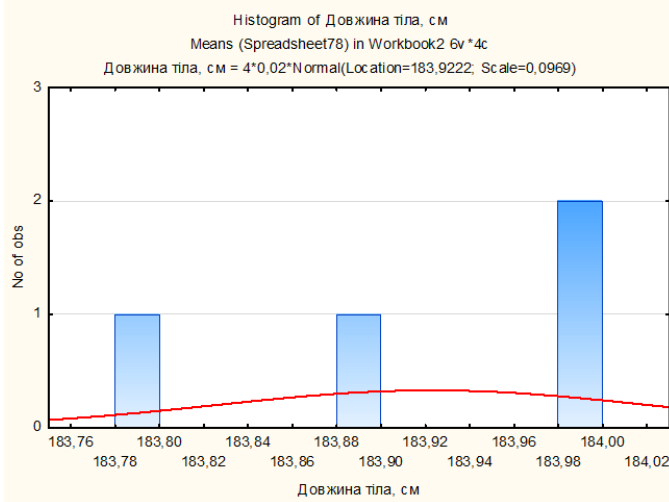


Рис. 131. Гістограми для візуалізації описової статистики (розподіл)

Багато інших процедур графічної візуалізації даних також доступні у діалоговому вікні **Описові статистики**. Ці опції описуються нижче.

**Діаграма розмаху.** У вкладці **All cases (Усі спостереження)** натисніть кнопку **Box plot of means (Діаграма розмаху)** для отримання діаграми розмаху для незалежних змінних (рис. 132). Першим відобразиться стандартне діалогове вікно Вибір змінних, в якому потрібно вказати декілька або усі змінні, а потім натиснути кнопку **ОК**.



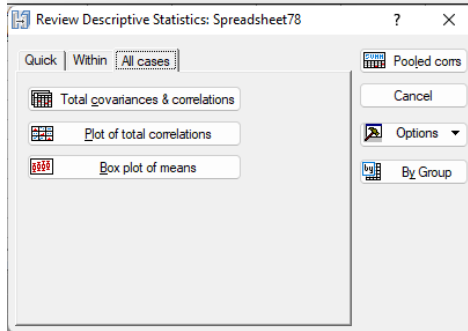


Рис. 132. Діалогове вікно *Review Descriptive Statistics* (Переглянути описові статистики)

Далі, відобразиться діалог **Тип діаграми розмаху**, в ньому відзначте опцію **Середнє/SD/1.96\*SD** і натисніть кнопку **ОК** (рис. 133).

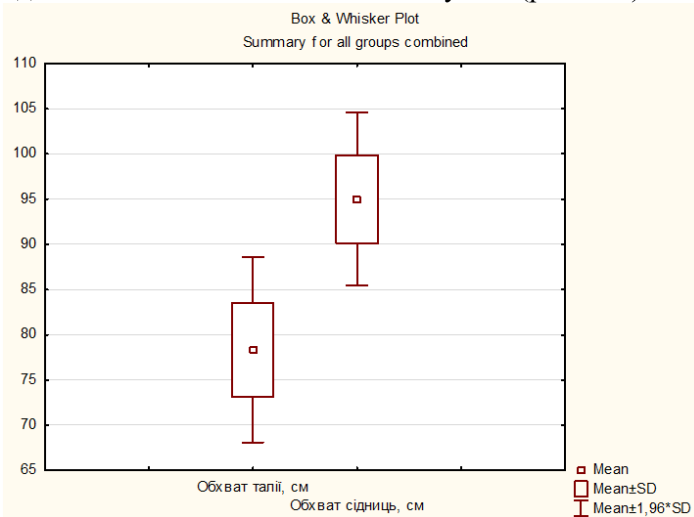


Рис. 133. Діаграма розмаху

Цей графік корисний для представлення розподілів змінних, які він описує за допомогою наступних трьох компонентів:

*Центральна точка* або лінія вказує на положення центральної області (тобто на середнє чи медіану);

*Прямокутник* вказує на характер мінливості навколо центрального положення (тобто квартилі, стандартні помилки або стандартні відхилення);

*Відрізки навколо прямокутників* вказують на діапазон значень змінної [наприклад, розмахи, стандартні відхилення, помножені на 1,96 стандартні відхилення (95% довірчий інтервал), помножені на 1,96 стандартні похибки середнього (95% довірчий інтервал)].

Ви можете подивитися на розподіл змінних усередині кожної групи, натиснувши кнопку **Box plot of means by group** (Діаграма розмаху за групами) у вкладці **Within** (Внутрішньогрупові статистики) та вибравши змінну **Довжина передпліччя** (рис. 134).

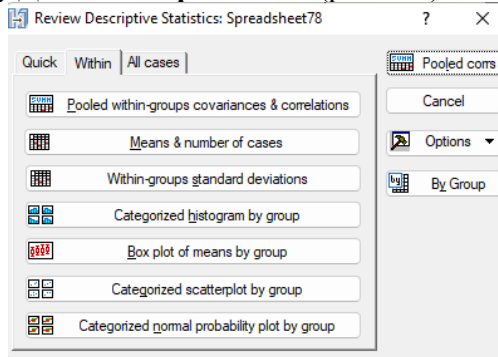


Рис. 134. Вкладка *Within* (Внутрішньогрупові статистики)

Тоді в наступному діалоговому вікні **Тип діаграми розмаху** виберіть команду **Середнє/SD/1.96\*SD** для вибору типу діаграми розмаху (рис. 135).

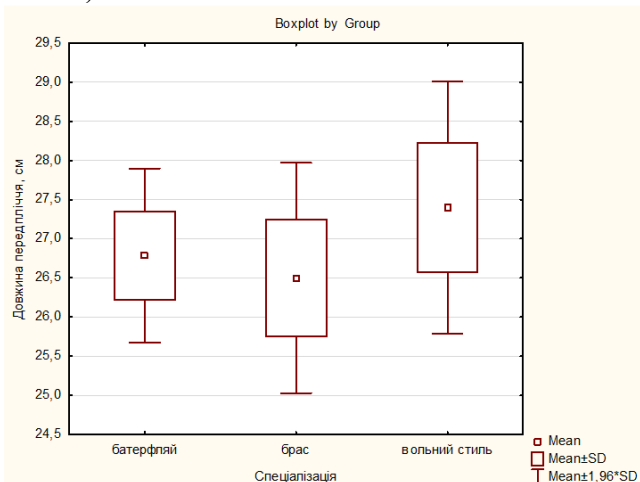


Рис. 135. Діаграма розподілу змінних усередині кожної групи за змінною *Довжина тіла*

**Категоризовані гістограми.** Ви можете графічно відобразити гістограми для змінних на кожному рівні групуючої змінної, натиснувши кнопку **Categorized histogram by group (Категоризована гістограма (за групами))** у вкладці **Within (Внутрішньогрупові статистики)** діалогового вікна **Review Descriptive Statistics (Описові статистики)**. Коли ви натискаєте на цю кнопку, ви можете вибрати змінні зі списку попередньо відібраних незалежних змінних. Для цього прикладу оберіть змінну **Висота стопи**. Гістограми для кожної із сукупностей, визначених на стартовій панелі, представлені нижче (рис. 136).

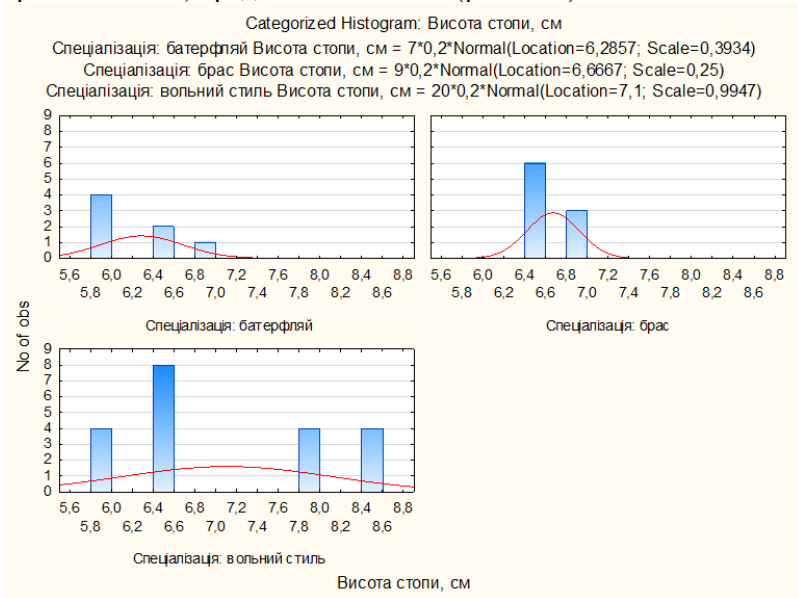


Рис. 136. Категоризовані гістограми за змінною *Висота стопи*

**Діаграма розсіювання.** Іншим цікавим типом графіків є діаграми розсіювання кореляцій між змінними, які беруть участь у аналізі. Для графічного огляду кореляцій між усіма змінними за допомогою діаграм розсіювання натисніть кнопку **Plot of total correlations (Графік повних кореляцій)** у вкладці **All cases (Усі спостереження)** діалогового вікна **Review Descriptive Statistics (Описові статистики)**. Виберіть усі спостереження у діалоговому вікні **Вибрати змінні** (рис. 137).

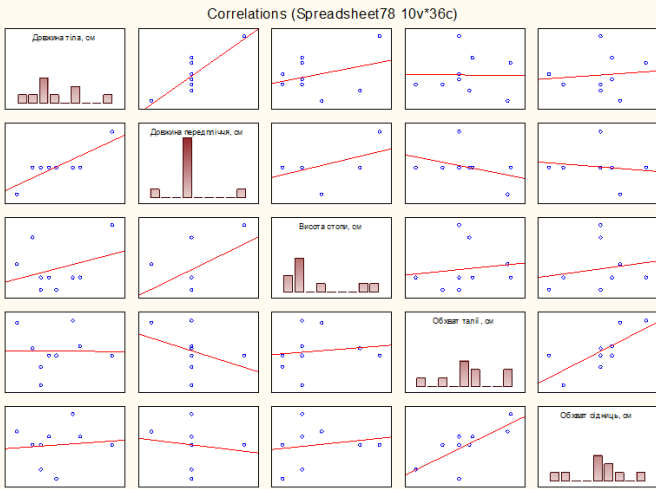


Рис. 137. Діаграми розсіювання кореляцій між змінними, які беруть участь у аналізі

**Вибір аналізу дискримінантної функції.** Тепер повернемося до первинної мети нашого аналізу. Натисніть кнопку **Cancel (Скасувати)** у діалоговому вікні **Descriptive Statistics (Описові статистики)**, щоб повернутися до діалогового вікна **Model Definition (Визначення Моделі)**. Щоб побачити, що відбувається кожному кроці дискримінантного аналізу, необхідно виконати покроковий аналіз. У вкладці **Advanced (Додатково)**, у списку **Method (Метод)** встановіть **Forward stepwise (Покроковий з включенням)**. При такій установці програма вводитиме змінні в модель поступово, одну за одну, щоразу вибираючи змінну, що робить найбільший внесок у дискримінацію (рис. 138).

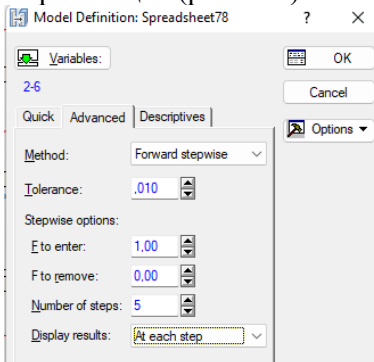


Рис. 138. Встановлення параметрів в діалоговому вікні *Model Definition (Визначення Моделі)*

**Правила зупинення.** Statistica перебуватиме в покроковому режимі доти, доки не відбудеться одна з чотирьох подій. Програма припинить покрокову процедуру, якщо:

Усі змінні введені або відкинуті, або

Досягнуто максимальну кількість кроків, встановлену в полі **Число кроків**, або

Немає інших змінних поза моделлю, що мають більше значення статистики F, ніж значення F-включити, зазначене в цьому діалоговому вікні, і коли в моделі немає інших змінних, що мають менше значення F, ніж значення F-виключити, зазначене в цьому діалоговому вікні, або

Будь-яка змінна на наступному кроці має значення толерантності менше, ніж обране значення Толерантність.

**F для увімкнення/виключення.** При покроковому аналізі з включенням програма відбирає змінні включення, що дають найбільш значущий єдиний (додатковий) внесок у дискримінацію між сукупностями; тобто програма вибирає змінні з найбільшим значенням F (більшим, ніж відповідне вказане користувачем значення F-включити). При виконанні кроків з виключенням програма буде відбирати для виключення найменш значущі змінні, тобто змінні з найменшим значенням F (меншим, ніж відповідне вказане користувачем значення F-виключити). Тому, якщо ви хочете ввести всі змінні в покроковий аналіз із включенням, покладіть значення F-включити настільки малим, наскільки це можливо (а значення F-виключити дорівняйте до нуля).

Якщо ви бажаєте виключити всі змінні з моделі одну за іншою, то виберіть значення F-включити дуже великим (наприклад, 9999), і надайте значення F-виключити також дуже велике, але трохи менше значення, ніж значення F-включити (наприклад, 9998). Згадаймо, що значення F-включити повинно мати завжди більше значення, ніж значення F-виключити.

**Толерантність.** Сенс поняття Толерантності полягає у тому, що на кожному кроці програма обчислює для кожної змінної множинну кореляцію (R-квадрат) з усіма іншими змінними, включеними до моделі. Значення толерантності змінної визначається як 1 мінус R-квадрат. Тому значення толерантності є мірою надмірності змінної.

Наприклад, якщо змінна, призначена для включення в модель, має значення толерантності, що дорівнює 0,01, то ця змінна може розглядатися як на 99% надлишкова вже з включеними змінними. Зазначимо, що коли одна або більше включених змінних стають надмірними, то матриця дисперсій/коваріацій для змінних, включених в

модель, може виявитися незворотною, і дискримінантний аналіз не зможе бути виконаний.

Загалом рекомендується залишати значення толерантності, що дорівнює 0.01 і встановлюється у програмі за промовчанням. Якщо змінна включена в модель і скоротна з іншими змінними більш ніж на 99%, її практичний внесок у поліпшення якості дискримінації дуже незначний. Більш важливо, якщо ви покладете для толерантності значно менше значення, то помилки округлення можуть призвести до нестійких результатів.

**Початок аналізу.** Після перегляду різних параметрів у цьому діалоговому вікні можна продовжити роботу звичайним чином, тобто не змінюючи жодних установок, прийнятих за замовчуванням. Однак, щоб переглянути результати на кожному кроці, встановіть у полі **Display results (Виведення результату)** значення **At each step (На кожному кроці)**. Тепер натисніть кнопку ОК, щоб розпочати дискримінантний аналіз.

### *Перегляд результатів дискримінантного аналізу.*

Результати на кроці 0. Спочатку з'являються Результати дискримінантного аналізу на нульовому кроці. Слова Крок 0 означають, що ще жодної змінної моделі не було включено (рис. 139).

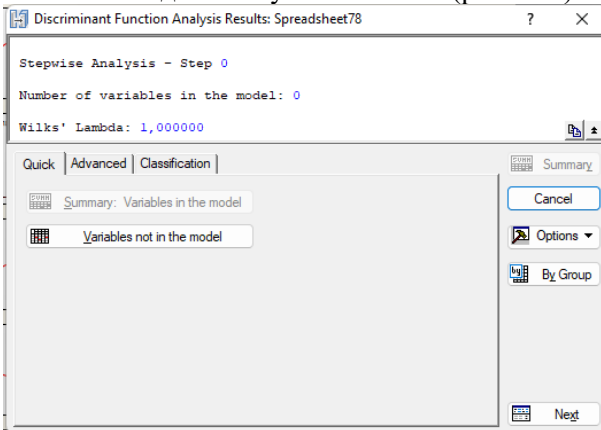


Рис. 139. Результати аналізу на кроці 0

Оскільки жодної змінної не було включено в модель, більшість операцій ще недоступно (і вони неактивні). Однак ви можете подивитися на змінні, які не включені в модель, натиснувши кнопку **Variables not in the model (Змінні поза моделлю)** (рис. 140).

N=36	Variables currently not in the model (Spreadsheet78)					
	Wilks' Lambda	Partial Lambda	F to enter	p-value	Toler.	1-Toler. (R-Sqr.)
Довжина тіла, см	0,999153	0,999153	0,01399	0,986108	1,000000	0,00
Довжина передпліччя, см	0,772577	0,772577	4,85711	0,014159	1,000000	0,00
Висота стопи, см	0,842857	0,842857	3,07627	0,059558	1,000000	0,00
Обхват талії, см	0,470860	0,470860	18,54230	0,000004	1,000000	0,00
Обхват сідниць, см	0,673769	0,673769	7,98909	0,001481	1,000000	0,00

Рис. 140. Параметри змінних, які не включені в модель, на кроці 0

**Лямбда Вілк (Wilks' Lambda).** Загалом, статистика Вілкса лямбда є стандартною статистикою, яка використовується у стандартній статистиці для позначення статистичної значущості потужності дискримінації у поточній моделі. Її значення змінюється від 1,0 (немає дискримінації) до 0,0 (повна дискримінація). Кожне значення у першій колонці таблиці, наведеної вище, є значенням статистики Вілкса лямбда після того, як відповідна змінна вводиться в модель.

**Приватна лямбда Вілкс.** Це статистика лямбда Вілкса для одиночного вкладу відповідної змінної дискримінації між сукупностями. Це значення можна розглядати як аналог приватного коефіцієнта кореляції. Оскільки лямбда з величиною 0,0 позначає повну дискримінацію, то чим нижче її значення в цьому стовпці, тим більше одиночний внесок відповідної змінної у ступінь дискримінації. Оскільки у модель ще не було введено жодної змінної, приватна лямбда Вілкса на кроці 0 дорівнює статистиці лямбда Вілкса після введення змінної (див. значення, що містяться в першій колонці таблиці).

**F-включити та р-значення.** Статистика Вілкса лямбда може бути перетворена до стандартного значення F, і ви можете обчислити відповідне р-значення для кожного значення F. Однак не слід в загальному випадку приймати ці р-значення в якості вирішального результату. Звичайно, не можна розраховувати на шанс, коли до аналізу включаються деякі змінні без володіння щодо них якими-небудь апріорними гіпотезами. Вибір для інтерпретації лише тих, щодо яких сподіваються, що вони будуть значущими, також є неприйнятним.

Отже, є велика різниця між апріорним передбаченням значущого впливу для окремої змінної і знаходженням потім цієї змінної, як це показує приклад із вибором серед 100 змінних однієї зі значним впливом. Не розглядаючи деталі, в суто практичних термінах зауважимо, що в останньому випадку не дуже схоже, що ви визначите ту саму змінну як значущу при повторенні дослідження. Коли звітують про результати аналізу дискримінантної функції, необхідно постаратися не залишити

враження, що спочатку вибрали лише значні змінні (виходячи з деяких теоретичних міркувань), тоді, коли вони були обрані оскільки вони добре «працюють».

Поглянувши на таблицю вище, можна побачити, що найбільші значення величини **F to enter (F-включити)** дає змінна Обхват талії. Таким чином, цю змінну буде введено в модель на наступному (першому) кроці.

**Толерантність та R-квадрат.** Значення толерантності обговорювалося і раніше у цьому розділі. Повторюючи це визначення, скажемо, що воно визначається як 1 мінус R-квадрат для відповідної змінної з усіма іншими змінними в моделі. Це значення толерантності дає інформацію про надмірність цієї змінної. Коли інші змінні ще обрані, все R-квадрат дорівнюють 1,0.

Результати на кроці 2. Тепер натисніть кнопку **Next (Далі)**, щоб перейти до наступного кроку. Крок 1 тут не буде розглядатися, тож натисніть знову на кнопку **Next (Далі)**, щоб перейти до кроку 2 (модель з двома змінними). Діалогове вікно Результати дискримінантного аналізу виглядатиме подібно до цього вікна на рис. 141:

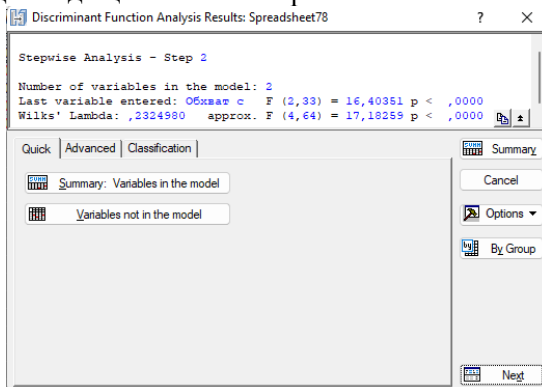


Рис. 141. Результати аналізу на кроці 2

Зовні дискримінація між видами спеціалізації значуща (статистика Вілкса лямбда = .232;  $F = 17.2$ ,  $p < 0.0001$ ). Погляньмо тепер на незалежні вклади кожної змінної моделі в передбачення.

**Змінні моделі.** Натисніть кнопку **Variables in the model (Змінні в моделі)**. З'явиться таблиця результатів для поточних змінних моделі. Як ви можете бачити, обидві змінні високо значущі (рис. 142).



Discriminant Function Analysis Summary (Spreadsheet78)						
Step 2, N of vars in model: 2; Grouping: Спеціалізація (3 grps)						
Wilks' Lambda: .23250 approx. F (4,64)=17,183 p< .0000						
N=36	Wilks' Lambda	Partial Lambda	F-remove (2,32)	p-value	Toler.	1-Toler. (R-Sqr.)
Обхват галії, см	0,673769	0,345071	30,36732	0,000000	0,332132	0,667868
Обхват сідниць, см	0,470860	0,493774	16,40351	0,000012	0,332132	0,667868

Рис. 142. Параметри змінних, які включені в модель, на кроці 2

**Змінні поза моделлю.** Натисніть на кнопку **Variables not in the model** (Змінні поза моделлю), щоб отримати таблицю з тими самими статистиками, що ми бачили раніше (рис. 143).

Variables currently not in the model (Spreadsheet78)						
Df for all F-tests: 2,31						
N=36	Wilks' Lambda	Partial Lambda	F to enter	p-value	Toler.	1-Toler. (R-Sqr.)
Довжина тіла, см	0,221544	0,952887	0,76636	0,473303	0,927939	0,072061
Довжина передпліччя, см	0,214671	0,923324	1,28718	0,290393	0,991746	0,008254
Висота стопи, см	0,114983	0,494556	15,84122	0,000018	0,568350	0,431650

Рис. 143. Параметри змінних, які не включені в модель, на кроці 2

Як ви бачите, 2 змінні, які ще поза моделлю, мають значення F-включити включити більші, ніж 1. Ви знаєте, що через це покрокова процедура буде продовжуватися, і наступна змінна, яка буде вводиться в модель – це змінна Висота стопи.

Результати на кроці 4 і 5 (останній крок). Натисніть кнопку **Next** (Далі) в діалоговому вікні Результати дискримінантного аналізу, щоб перейти до наступного кроку аналізу. Крок 3 не буде тут розглядатися, так що натискаємо знову на кнопку Далі, щоб перейти до фінального кроку в цьому аналізі – Крок 5 (показано вище) (рис. 144).

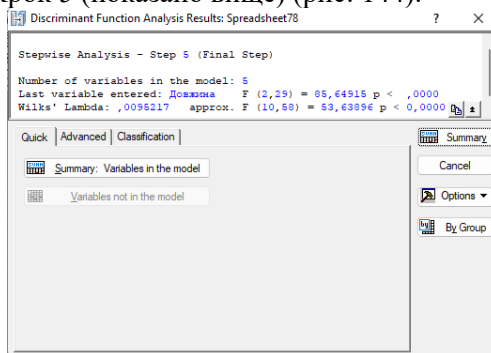


Рис. 144. Результати аналізу на кроці 5 (останньому)

Тепер натиснемо на кнопку **Variables in the model** (Змінні в моделі) для огляду незалежних вкладів кожної змінної в загальну дискримінацію між видами спеціалізації (рис. 145).

Discriminant Function Analysis Summary (Spreadsheet78)						
Step 5, N of vars in model: 5; Grouping: Спеціалізація (3 grps)						
Wilks' Lambda: .00952 approx. F (10,58)=53.639 p<0,0000						
N=36	Wilks' Lambda	Partial Lambda	F-remove (2,29)	p-value	Toler.	1-Toler. (R-Sqr.)
Обхват талії, см	0,072614	0,131128	96,0792	0,000000	0,056478	0,943522
Обхват сідниць, см	0,103072	0,092379	142,4618	0,000000	0,059450	0,940550
Висота стопи, см	0,015839	0,601155	9,6203	0,000624	0,283401	0,716599
Довжина тіла, см	0,094174	0,101107	128,9121	0,000000	0,011618	0,988382
Довжина передпліччя, см	0,065765	0,144784	85,6492	0,000000	0,013298	0,986702

Рис. 145. Параметри змінних, які включені в модель, на кроці 5 (останньому)

Часткова статистика Вілкса лямбда показує, що змінна Висота стопи дає вклад найбільше, змінна Довжина передпліччя – друга за значенням вкладу, змінна Обхват талії – третя за значенням вкладу, Довжина тіла – четверта за значенням вкладу і змінна Обхват сідниць вносить у загальну дискримінацію найменше. (Згадаймо, що чим менша статистика Вілкса лямбда, тим більший внесок у загальну дискримінацію.) Тому ви можете встановити на цій стадії дослідження, що Висота стопи є головною змінною, яка дозволяє здійснювати дискримінацію між різними видами спеціалізації в плаванні. Для отримання подальших результатів щодо природи дискримінації слід провести канонічний аналіз. Стартову панель Канонічного аналізу можна викликати відповідною кнопкою у вкладці **Advanced (Додатково)** діалогового вікна Результати дискримінантного аналізу.

**Канонічний аналіз.** Щоб побачити, як п'ять змінних поділяють різні сукупності (види спеціалізації), обчислимо дійсну дискримінантну функцію. Натисніть кнопку **Perform canonical analysis (Канонічний аналіз)** для виконання канонічного аналізу та відкрийте діалогове вікно **Canonical Analysis (Канонічний аналіз)** (рис. 146).

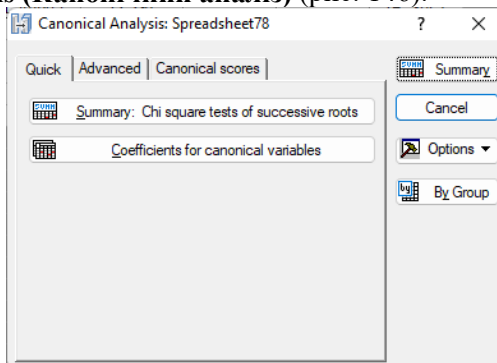


Рис. 146. Діалогове вікно Canonical Analysis (Канонічний аналіз)

Програма обчислить різні незалежні (ортогональні) дискримінаційні функції. Кожна подальша дискримінантна функція вноситиме все менший і менший внесок у загальну дискримінацію. Максимальна кількість функцій, що оцінюються, або дорівнює числу змінних або числу сукупностей мінус один, залежно від того, яке число менше. У нашому випадку оцінюються дві дискримінуючі функції.

**Значення коренів.** Спочатку визначимо, чи обидві дискримінантні функції (корінь) статистично значущими. Натисніть на кнопку **Chi-Square Tests of Successive Roots (Критерій Хі-квадрат послідовного кореня)** і побачите наступну таблицю (рис. 147):

Chi-Square Tests with Successive Roots Removed (Spreadsheet78)						
Roots Removed	Eigen-value	Canonicl R	Wilks' Lambda	Chi-Sqr.	df	p-value
0	60,57836	0,991847	0,009522	144,2797	10	0,000000
1	0,70553	0,643172	0,586329	16,5501	4	0,002363

Рис. 147. Таблиця *Chi-Square Tests of Successive Roots (Критерій Хі-квадрат послідовного кореня)*

Загалом, ця таблиця дає звіт про покроковий критерій із включенням для всіх канонічних коренів. Перший рядок завжди містить критерій значущості для всіх коренів, другий рядок дає дані про значущість коренів, що залишилися після видалення першого кореня і т.д. Таким чином, ця таблиця говорить вам, як багато канонічних коренів (дискримінуючих функцій) слід інтерпретувати. У цьому прикладі обидві дискримінантні (або канонічні) функції статистично значущі. Тому ви отримаєте два окремі висновки (інтерпретації), тому що різні антропометричні параметри дозволяють вам розрізняти види спеціалізації в плаванні.

**Коефіцієнти дискримінантної функції.** Натисніть кнопку **Coefficients for Canonical Variables (Коефіцієнти для канонічних змінних)** у діалоговому вікні Канонічний аналіз. Будуть отримані дві таблиці, одна для вихідних коефіцієнтів та інша для стандартизованих коефіцієнтів. Погляньмо тепер на Вихідні коефіцієнти (рис. 148).

Variable	Raw Coefficients (Spreads for Canonical Variables)	
	Root 1	Root 2
<b>Обхват талії, см</b>	<b>-1,0558</b>	<b>-0,2853</b>
Обхват сідниць, см	0,9338	0,3224
Висота стопи, см	1,3851	0,9939
Довжина тіла, см	-2,4761	0,3787
Довжина передпліччя, см	10,4701	-2,4726
Constant	156,5659	-17,8200
Eigenval	60,5784	0,7055
Cum.Prop	0,9885	1,0000

Рис. 148. Таблиця вихідних коефіцієнтів дискримінантної функції

Вихідні тут означає, що коефіцієнти тут можуть бути використані разом із спостереженими даними для обчислення (рядків) ваги дискримінантної функції. Стандартизовані коефіцієнти – це ті коефіцієнти, які зазвичай використовуються для інтерпретації, тому що вони відносяться до нормованих змінних і тому повинні знаходитись у порівняних масштабах (рис. 149).

Variable	Standardized Coefficients for Canonical Variables	
	Root 1	Root 2
<b>Обхват талії, см</b>	<b>-3,89515</b>	<b>-1,05265</b>
Обхват сідниць, см	3,84425	1,32724
Висота стопи, см	1,08442	0,77814
Довжина тіла, см	-8,82502	1,34991
Довжина передпліччя, см	7,99225	-1,88746
Eigenval	60,57836	0,70553
Cum.Prop	0,98849	1,00000

Рис. 149. Таблиця стандартизованих коефіцієнтів дискримінантної функції

Перша дискримінантна функція зважується найважче показниками довжини (змінні Довжина тіла та Довжина передпліччя). Інші три змінні також дають вклад у цю функцію. Друга функція, відзначена головним тими самими параметрами.

**Власні значення.** У таблиці вище наведено власні значення (корені) кожної дискримінантної функції і кумулятивна частка поясненої дисперсії, накопиченої кожною функцією. Як ви можете бачити, перша функція відповідальна за 99% поясненої дисперсії, тобто 99% всієї дискримінаційної потужності пояснюється цією функцією. Таким чином, ясно, що ця перша функція найбільш "важлива".

**Коефіцієнти факторної структури.** Ці коефіцієнти (які можуть бути виведені за допомогою кнопки **Factor structure (Факторна структура)** у вкладці **Advanced (Додатково)** діалогового вікна Канонічний аналіз

представляють кореляцію між змінними функцією дискримінації і в загальному випадку використовуються для інтерпретації "значущості" дискримінуючої функції.

У дослідженнях у сфері освіти чи психології іноді бажають надати функцій осмислені назви (тобто. "зовнішні версії", "мотивації призначення"), використовуючи самі міркування, що у факторному аналізі. У цих випадках інтерпретація факторів має бути заснована на коефіцієнтах факторної структури. Однак таку інтерпретацію не можна розглянути в цьому прикладі (рис. 150).

		Factor Structure Matrix (Spreadsheet78) Correlations Variables - Canonical Roots (Pooled-within-groups correlations)	
Variable	Root 1	Root 2	
Обхват талії, см	-0,126281	0,472835	
Обхват сідниць, см	-0,030884	0,777419	
Висота стопи, см	0,051584	0,189178	
Довжина тіла, см	-0,003741	0,000796	
Довжина передпліччя, см	0,067302	-0,168283	

Рис. 150. Коефіцієнти факторної структури

**Середні канонічні змінні.** Ви тепер знаєте, які змінні беруть участь у дискримінації між різними видами спеціалізації в плаванні. Наступне завдання полягає в тому, щоб визначити природу дискримінації для кожного канонічного кореня. На першому кроці поглянемо на канонічні середні. Натисніть кнопку **Means of canonical (Середні канонічних змінних)** у вкладці **Advanced (Додатково)** діалогового вікна Канонічний аналіз (рис. 151).

		Means of Canonical Variables	
Group	Root 1	Root 2	
батарфляй	-8,62011	-1,34681	
брас	-8,10300	1,08421	
вольний стиль	6,66339	-0,01651	

Рис. 151. Канонічні середні за спеціалізаціями

Очевидно, що перша дискримінантна функція відокремлює переважно спеціалізації «батарфляй» і «брас» від «вольного стилю». Канонічне середнє спеціалізації «батарфляй» і «брас» сильно відрізняється від канонічного середнього спеціалізації «вольний стиль». Щодо другої дискримінантної функції, ґрунтуючись на розглянутих раніше власних значеннях, її якість тепер набагато менша.

**Діаграма розсіювання для значень.** Швидкий спосіб візуалізації цих результатів полягає у виведенні на екран діаграми розсіювання двох дискримінантних функцій. У вкладці **Canonical scores (Канонічні значення)** діалогового вікна Канонічний аналіз натисніть кнопку **Scatterplot for canonical scores (Діаграма розсіювання для канонічних значень)** для відображення ненормованих значень на діаграмі Корінь 1 від Кореня 2 (рис. 152).

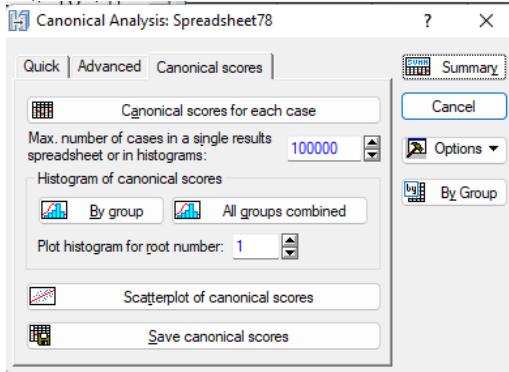


Рис. 152. Діалогове вікно для вибору діаграми розсіювання двох дискримінантних функцій

Ця діаграма (див. нижче) підтверджує таку інтерпретацію (рис. 153).

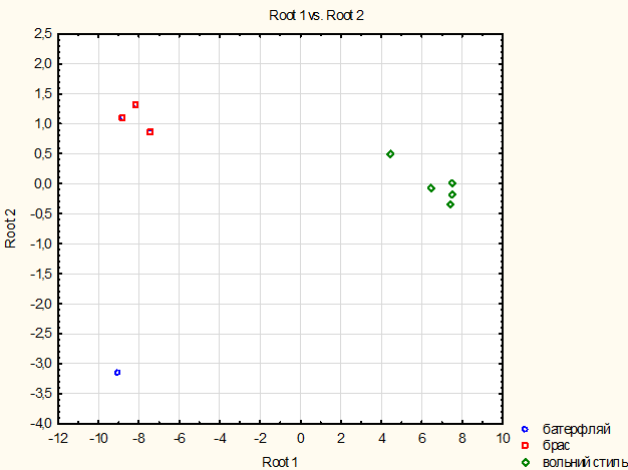


Рис. 153. Діаграма розсіювання двох дискримінантних функцій

Видно, що спортсмени спеціалізації вольний стиль представлені на діаграмі крапками далеко праворуч. Тому перша дискримінантна функція головним чином дискримінує між видом спеціалізації та двома іншими.

Друга функція, мабуть, дає деяку дискримінацію між спортсменами спеціалізації батерфляй (які мають негативні значення для другої канонічної функції) та іншою (з переважно позитивними значеннями). Проте дискримінація тут не настільки чітка, як це має місце для першої канонічної функції (кореня).

**Підсумок.** Для підбиття підсумків зауважимо, що найбільш значуща та ясна дискримінація можлива для плавців спеціалізацій «батерфляй» і «брас» з використанням першої функції дискримінації. Ця функція відзначена негативними коефіцієнтами для обхвату талії та довжини тіла та позитивними вагами для висоти стопи, довжини передпліччя та обхвату сідниць.

**Класифікація.** Повернімося до діалогового вікна Результати аналізу дискримінантних функцій (натисніть кнопку Скасувати у діалоговому вікні Канонічний аналіз) та поверніться до завдання класифікації. Однією з цілей аналізу дискримінантної функції – дати досліднику можливість провести класифікацію об'єктів. Тепер подивимося, як добре побудовані дискримінантні функції класифікують спортсменів.

**Функції класифікації.** Погляньмо спочатку функції класифікації. Не повинно виникнути проблем із застосуванням дискримінантних функцій. Функції класифікації обчислюються для кожної сукупності і можуть безпосередньо застосовуватися для класифікації об'єктів. Можна класифікувати спостереження в ту сукупність, для якої обчислено найбільшу класифікаційну вагу. Натисніть кнопку **Classification functions (Опції класифікації)** у вкладці **Classification (Класифікація)** діалогового вікна Результати аналізу дискримінантних функцій, щоб побачити ці функції (рис. 154).

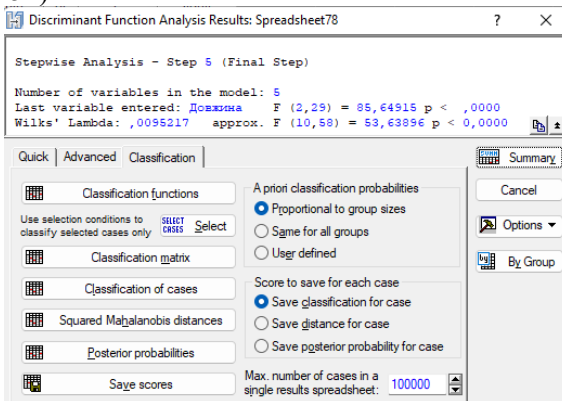


Рис. 154. Діалогове вікно Результати аналізу дискримінантних функцій

Ви можете використовувати ці функції, щоб визначити перетворення для нових змінних. Коли ви введете нове спостереження, програма автоматично обчислить класифікаційну вагу для кожної сукупності (рис. 155).

Variable	Classification Functions: grouping: Спеціалізація		
	батерфляй p=,19444	брас p=,25000	вольний стиль p=,55556
Обхват галії, см	168,0	166,8	151,5
Обхват сідниць, см	-148,1	-146,9	-133,4
Висота стопи, см	-216,1	-213,0	-193,6
Довжина тіла, см	430,9	430,6	393,6
Довжина передпліччя, см	-1776,3	-1776,9	-1619,5
Constant	-15102,8	-15060,3	-12716,7

Рис. 155. Таблиця із функціями класифікації

**Апріорні ймовірності.** Можна задати різні апріорні ймовірності для кожної сукупності (виберіть команду Задані користувачем у полі Апріорні ймовірності класифікації у вкладці Класифікація діалогового вікна Результати аналізу дискримінантних функцій). Це ймовірність того, що спостереження належить відповідній сукупності без використання будь-якої інформації про значення змінних у моделі. Апріорні ймовірності можуть значно підвищити точність класифікації. Ви також можете обмежитися обчисленням результатів для відібраних спостережень (натисніть кнопку Вибрати). Це корисно, зокрема, якщо ви хочете провести оцінку якості результатів аналізу дискримінантної функції з новими даними. Для нових спостережень, однак, тепер приймається установка за пропорційними розмірами груп.

**Таблиця класифікації.** Натисніть кнопку **Classification matrix (Матриця класифікації)**. У таблиці результатів, друга лінія в заголовку кожної колонки наводить апріорні ймовірності класифікації (рис. 156).

Classification Matrix (Spreadsheet78)				
Rows: Observed classifications				
Columns: Predicted classifications				
Group	Percent Correct	батерфляй p=,19444	брас p=,25000	вольний стиль p=,55556
батерфляй	57,1429	4	3	0
брас	100,0000	0	9	0
вольний стиль	100,0000	0	0	20
Total	91,6667	4	12	20

Рис. 156. Таблиця класифікації

Оскільки є різна кількість спортсменів різних спеціалізацій, ви обираєте ці можливості пропорційно обсягам вибірок, апріорні ймовірності для кожної сукупності різні. У першому стовпці таблиці ви бачите відсоток спостережень, які правильно класифіковані кожної



сукупності отриманими функціями класифікації. Стовпці, що залишилися, дають число випадків правильної і неправильної класифікації для кожної сукупності.

**Апріорна у порівнянні з апостеріорною класифікацією.** Коли ви класифікуєте спостереження, за якими вже була знайдена функція, що дискримінує, ви зазвичай отримуєте порівняно хорошу дискримінацію (хоча зазвичай не настільки хорошу, як у цьому прикладі). Однак цю класифікацію розглядають лише як діагностичний засіб ідентифікації сильних і слабких сторін отриманих дискримінантних функцій, оскільки ці класифікації не є апріорними передбаченнями, а скоріше апостеріорними. Тільки якщо класифікуються різні (нові) зразки, цю таблицю можна інтерпретувати термінах потужності дискримінації. Тому було б несправедливо заявляти, що ви можете успішно дискримінувати спортсменів трьох спеціалізацій на 98% у всіх випадках, ґрунтуючись лише на п'яти вимірах. Оскільки ви розраховуєте на удачу, то можете очікувати значно меншу точність при класифікації нових зразків (спортсменів).

#### **Класифікація спостережень.**

**Відстань Махаланобіса та апостеріорні ймовірності.** Тепер повернемося знову до діалогового вікна Результати аналізу дискримінантних функцій. Спостереження класифікуються в сукупності, до яких вони ближчі, ніж до інших. Відстань Махаланобіса є мірою відстані, яку можна використовувати в багатовимірному просторі, визначеному змінними моделями. Ви можете обчислити відстань між спостереженням та центром кожної сукупності (тобто центроїдом сукупності, визначеного відповідним середнім сукупністю для кожної змінної). Що ближче спостереження до центроїду групи, то більшою мірою ви можете бути впевнені, що це спостереження належить цій групі. Відстань Махаланобіса може бути отримана шляхом натискання на кнопку **Squared Mahalanobis distances (Квадрати відстані Махаланобіса)** у вкладці **Classification (Класифікація)** (рис. 157).

Squared Mahalanobis Distances from Group Centroids (Spreadsheet78) Incorrect classifications are marked with *				
Case	Observed Classif.	бaтepфлaй p=.19444	бpaс p=.25000	вoльний cтиль p=.55556
* 1	бaтepфлaй	9,4574	3,6905	202,3264
* 2	бaтepфлaй	9,2293	3,6652	242,8501
* 3	бaтepфлaй	10,2423	3,0415	222,9937
4	бaтepфлaй	3,4747	18,9179	255,5358
5	бaтepфлaй	3,4747	18,9179	255,5358
6	бaтepфлaй	3,4747	18,9179	255,5358
7	бaтepфлaй	3,4747	18,9179	255,5358
8	бpaс	9,4574	3,6905	202,3264
9	бpaс	9,2293	3,6652	242,8501
10	бpaс	10,2423	3,0415	222,9937
11	бpaс	9,4574	3,6905	202,3264
12	бpaс	9,2293	3,6652	242,8501
13	бpaс	10,2423	3,0415	222,9937
14	бpaс	9,4574	3,6905	202,3264
15	бpaс	9,2293	3,6652	242,8501
16	бpaс	10,2423	3,0415	222,9937
17	вoльний cтиль	264,4701	248,2368	3,4605
18	вoльний cтиль	174,5908	158,3149	6,0539
19	вoльний cтиль	264,3177	247,2309	3,8174
20	вoльний cтиль	234,6720	219,0751	5,4779
21	вoльний cтиль	261,4701	246,1759	3,9355
22	вoльний cтиль	264,4701	248,2368	3,4605
23	вoльний cтиль	174,5908	158,3149	6,0539
24	вoльний cтиль	264,3177	247,2309	3,8174
25	вoльний cтиль	234,6720	219,0751	5,4779
26	вoльний cтиль	261,4701	246,1759	3,9355
27	вoльний cтиль	264,4701	248,2368	3,4605
28	вoльний cтиль	174,5908	158,3149	6,0539
29	вoльний cтиль	264,3177	247,2309	3,8174
30	вoльний cтиль	234,6720	219,0751	5,4779
31	вoльний cтиль	261,4701	246,1759	3,9355
32	вoльний cтиль	264,4701	248,2368	3,4605
33	вoльний cтиль	174,5908	158,3149	6,0539
34	вoльний cтиль	264,3177	247,2309	3,8174
35	вoльний cтиль	234,6720	219,0751	5,4779
36	вoльний cтиль	261,4701	246,1759	3,9355

Рис. 157. Відстань Махаланобіса

Ви можете безпосередньо обчислити ймовірність того, що спостереження належить певній сукупності. Це умовна ймовірність у тому сенсі, що вона залежить від вашого знання значень змінних у моделі. Тому ці ймовірності називаються апостеріорними ймовірностями. Ви можете отримати їх, натиснувши кнопку Апостеріорні ймовірності. Зауважимо, що, як і у разі матриць класифікації, ви можете відібрати зразки для класифікації та вказати різні апріорні ймовірності

**Конкретні класифікації.** Нижче показано конкретну класифікацію спостережень (спортсменів; кнопка **Posterior probabilities (Класифікація спостережень)**) (рис. 158).

Posterior Probabilities (Spreadsheet78)				
Incorrect classifications are marked with *				
Case	Observed Classif.	батарфляй p= .19444	брас p= .25000	вольний стиль p= .55556
* 1	батарфляй	0.041697	0.958303	0.000000
* 2	батарфляй	0.045940	0.954060	0.000000
* 3	батарфляй	0.020801	0.979199	0.000000
4	батарфляй	0.999431	0.000569	0.000000
5	батарфляй	0.999431	0.000569	0.000000
6	батарфляй	0.999431	0.000569	0.000000
7	батарфляй	0.999431	0.000569	0.000000
8	брас	0.041697	0.958303	0.000000
9	брас	0.045940	0.954060	0.000000
10	брас	0.020801	0.979199	0.000000
11	брас	0.041697	0.958303	0.000000
12	брас	0.045940	0.954060	0.000000
13	брас	0.020801	0.979199	0.000000
14	брас	0.041697	0.958303	0.000000
15	брас	0.045940	0.954060	0.000000
16	брас	0.020801	0.979199	0.000000
17	вольний стиль	0.000000	0.000000	1.000000
18	вольний стиль	0.000000	0.000000	1.000000
19	вольний стиль	0.000000	0.000000	1.000000
20	вольний стиль	0.000000	0.000000	1.000000
21	вольний стиль	0.000000	0.000000	1.000000
22	вольний стиль	0.000000	0.000000	1.000000
23	вольний стиль	0.000000	0.000000	1.000000
24	вольний стиль	0.000000	0.000000	1.000000
25	вольний стиль	0.000000	0.000000	1.000000
26	вольний стиль	0.000000	0.000000	1.000000
27	вольний стиль	0.000000	0.000000	1.000000
28	вольний стиль	0.000000	0.000000	1.000000
29	вольний стиль	0.000000	0.000000	1.000000
30	вольний стиль	0.000000	0.000000	1.000000
31	вольний стиль	0.000000	0.000000	1.000000
32	вольний стиль	0.000000	0.000000	1.000000
33	вольний стиль	0.000000	0.000000	1.000000
34	вольний стиль	0.000000	0.000000	1.000000
35	вольний стиль	0.000000	0.000000	1.000000
36	вольний стиль	0.000000	0.000000	1.000000

*Рис. 158. Апостеріорні ймовірності*

Класифікація проводиться на перший, другий та третій вибори. Стовець із заголовком 1 містить перший вибір класифікації, тобто код сукупності, для якої відповідний зразок має найвищу апостеріорну ймовірність. Рядки, позначені зірочкою (\*), вказують на неправильно класифіковані зразки. Знову в цьому прикладі точність класифікації дуже висока, навіть з огляду на те, що це апостеріорна класифікація. Така точність рідко досягається у соціальних науках.

**Підсумок.** Цей приклад ілюструє основну ідею аналізу дискримінантних функцій. У цілому, у багатьох випадках, ця техніка може бути застосовна для реальних сукупностей. Однак, як зазначено у багатьох місцях, якщо вашою метою є коректна класифікація невідомих зразків, тоді слід провести дослідження у два етапи: спочатку побудувати функції класифікації та потім провести оцінку їх якості.

### Практична частина

Група школярів пройшла тестування для визначення рівня координаційних здібностей. За результатами тестування визначено рівень підготовленості: високий, середній або низький.

Наскільки суттєвим є розподіл за рівнями підготовленості за результатами проведеного тестування? На основі отриманих даних знайдіть дискримінантні рівняння. Визначте, хто з досліджуваних школярів має бути віднесений до іншого рівня підготовленості.

Проба Ромберга, с	Стрибок з поворотом на 360 град., град.	Три перекиди вперед, с	Човниковий біг 3x10 м, с	Рівень
60	690	2,5	8	високий
60	680	2,5	8	високий
54	560	2,7	8,1	високий
40	540	2,8	8,2	високий
32	540	2,9	8,2	високий
27	540	2,9	8,2	високий
27	520	3	8,3	високий
27	460	3	8,4	високий
26	440	3	8,6	високий
26	420	3	8,6	середній
25	420	3	8,6	середній
24	420	3,1	8,7	середній
24	410	3,1	8,7	середній
23	390	3,1	8,8	середній
19	380	3,2	8,8	середній
19	380	3,2	8,9	середній
18	360	3,2	9	середній
18	360	3,2	9	середній
16	360	3,2	9	середній
15	360	3,2	9	середній
15	360	3,2	9	середній
15	360	3,4	9,2	низький
14	360	3,4	9,4	низький
12	360	3,5	9,4	низький
11	360	3,6	9,4	низький
11	360	3,6	9,5	низький
10	360	3,8	9,6	низький
10	340	3,8	9,9	низький
10	330	3,8	11,2	низький
5	330	3,8	11,6	низький

**Список рекомендованої літератури**

1. Азарова А. О. Інформатика та комп'ютерна техніка (Частина 1) : навчальний посібник Вінниця : ВНТУ, 2012. 361 с.
2. Ахметов Р. Ф. Спортивна метрологія : навчальний посібник. Житомир: Вид-во ФОП Євенок О. О., 2017. 176 с.
3. Ашанін В. С., Жерновнікова Я. В., Долгополова Н. В. Формування науково-дослідної компетенції при вивченні дисципліни «Системно-інформаційні основи наукових досліджень у фізичній культурі та спорті». *Проблеми та перспективні напрями розвитку сучасного спорту: актуальні питання теорії та практики: матеріали Всеукр. наук.-практ. конф. (м. Харків, 5 квітня 2024 р.)*. Харків, 2024. С. 150–153.
4. Ашанін В. С., Жерновнікова Я. В., Пятисоцька С. С. Інформатика [навчальний посібник для студентів I-го курсу спеціальності 076 «Підприємництво, торгівля та біржова діяльність»]. Харків : ХДАФК, 2021. 148 с.
5. Ашанін В. С., Жерновнікова Я. В., Пятисоцька С. С. Комп'ютерна обробка даних експериментальних досліджень : навч. посіб. Харків : ХДАФК, 2024. 116 с.
6. Ашанін В. С., Пятисоцька С. С. Науково-методичні засади вивчення методів математико-статистичного аналізу даних у закладах вищої освіти фізкультурно-спортивного профілю. *Фізична культура, спорт і здоров'я : стан, проблеми та перспективи: матеріали ХХІІІ Міжнар. наук.-практ. конф. (м. Харків, 6 груд. 2023 р.)*. Харків, 2023. С. 177–178.
7. Ашанін В. С., Пятисоцька С. С. Системно-інформаційні основи наукових досліджень в фізичному вихованні та спорті : навчальний посібник. Харків : ХДАФК, 2019. 78 с.
8. Ашанін В. С., Пятисоцька С. С., Жерновнікова Я. В. Статистична обробка та аналіз інформації [навчальний посібник для здобувачів вищої освіти спеціальності 014 Середня освіта (Фізична культура)]. Харків : ХДАФК, 2022. 112 с.
9. Бахрушин В. Є. Методи аналізу даних : навчальний посібник для студентів. Запоріжжя : КПУ, 2011. 268 с.
10. Грицюк П. М., Остапчук О. П. Аналіз даних : навчальний посібник. Рівне : НУВГП, 2008. 218 с.
11. Донченко В. С., Сидоров М. В. Теорія ймовірностей та математична статистика для соціальних наук : навч. посіб. К. : ВПЦ «Київський університет», 2015. 400 с.

12. Жерновнікова Я. В. Формування змісту курсу «Статистична обробка та аналіз інформації» для майбутніх вчителів фізичної культури. *Proceedings of the scientific and pedagogical internship. Directions of improving the professional training of specialists in physical culture and sports.* Włocławek, 2023. P. 86-89.

13. Жерновнікова Я. В., Алексеєва І. А., Алексєнко Я. В. Використання електронних таблиць Microsoft Excel для обробки статистичних даних в галузі фізичного виховання. *Фізична культура, спорт і здоров'я : стан, проблеми та перспективи* : матеріали XXIII Міжнар. наук.-практ. конф. (м. Харків, 6 груд. 2023 р.). Харків, 2023. С. 177–178.

14. Кушлик-Дивульська О. І., Поліщук Н. В., Орел Б. П., Штабалуєк П. І. Теорія ймовірностей та математична статистика : навчальний посібник. К. : НТУУ «КПІ», 2014. 212 с.

15. Маханець Л. Л., Вінничук О. Ю., Григорків М. В. Статистика: лабораторний практикум у STATISTICA 12: навч. посіб. Чернівці : Чернівець. нац. ун-т ім. Ю. Федьковича, 2023. 161 с.

16. Нелюбов В. О., Куруца О. С. Основи інформатики. Microsoft Excel 2016 : навчальний посібник. Ужгород : ДВНЗ «УжНУ», 2018. 58 с.

17. Пятисоцька С. С., Жерновнікова Я. В. Інформаційні технології у фізичній культурі та спорті [методичні рекомендації для студентів I-го курсу спеціальності 017 Фізична культура і спорт]. Харків : ХДАФК, 2022. 52 с.

18. Руденко В. М. Математична статистика : навчальний посібник. К. : Центр учбової літератури, 2012. 304 с.

19. Телейко А. Б., Чорней Р. К. Матиматико-статистичні методи в соціології та психології : навчальний посібник. К. : МАУП, 2007. 424 с.

20. Фетісов В. С. Пакет статистичного аналізу даних STATISTICA : навч. посіб. Ніжин : НДУ ім. М. Гоголя, 2018. 114 с.



Навчальне видання

**Ашанін Володимир Семенович**

завідувач кафедри інформатики та біомеханіки Харківської державної академії фізичної культури, професор, кандидат фізико-математичних наук

**Пятисоцька Світлана Сергіївна**

доцент кафедри інформатики та біомеханіки Харківської державної академії фізичної культури, кандидат наук з фізичного виховання та спорту, доцент

**Жерновнікова Яна Вікторівна**

доцент кафедри інформатики та біомеханіки Харківської державної академії фізичної культури, кандидат педагогічних наук, доцент

**Петренко Юлія Іванівна**

доцент кафедри інформатики та біомеханіки Харківської державної академії фізичної культури, кандидат педагогічних наук

# **Системно-інформаційні основи наукових досліджень у фізичній культурі та спорті**

**Методичні рекомендації**

Харківська державна академія фізичної культури  
вул. Клочківська, 99, Харків, 61058