Ніжинський державний університет
імені Миколи Гоголя

В. С. Фетісов

Пакет статистичного аналізу даних STATISTICA

Навчальний посібник

Ніжин
2018

УДК 681.3.068 519.67
Ф45

Рекомендовано Вченою радою
Ніжинського державного університету імені Миколи Гоголя
(НДУ ім. М. Гоголя)
Протокол № 4 від 30.11.2017 р.

Рецензенти:

Чернишова Е. О. **–** доцент кафедри інформаційних технологій і аналізу даних Ніжинського державного університету ім. М. Гоголя, кандидат технічних наук;

Казачков І. В. – професор кафедри прикладної математики, інформатики та освітніх вимірювань Ніжинського державного університету ім. М. Гоголя, доктор технічних наук.

Фетісов В. С.

Ф45 Пакет статистичного аналізу даних STATISTICA : навч. посіб. / В. С. Фетісов. – Ніжин : НДУ ім. М. Гоголя, 2018. – 114 с.

АНОТАЦІЯ

Посібник містить опис популярного статистичного пакету STATISTICA, лабораторні завдання для самостійного виконання на комп'ютері, а також теоретичні відомості деяких статистичних методів.

Він буде корисним усім студентам, аспірантам і викладачам, які під час навчання, роботи або практичних досліджень здійснюють статистичний аналіз даних.

УДК 681.3.068 519.67

Загальні положення

Пакет *STATISTICA* – це універсальний пакет статистичного аналізу, в якому реалізовані основні математичні методи аналізу даних. Розробником пакету є фірма *StatSoft, Inc* (США). У 2014 р. ця фірма була поглинута корпорацією *Dell*, яка включила пакет *STATISTICA* до складу власної лінійки програмного забезпечення проблематики великих даних.

STATISTICA дозволяє проводити різні процедури (модулі) обробки статистичних даних (в термінології програми – *аналізи*):

1. Розрахунок описових статистик.
2. Аналіз динамічних рядів й прогнозування.
3. Множинна регресія.
4. Дискримінантний аналіз.
5. Аналіз відповідностей.
6. Кластерний аналіз.
7. Факторний аналіз.
8. Дисперсійний аналіз і та ін.

Крім загальних статистичних і графічних засобів *STATISTICA* має спеціалізовані модулі: для проведення соціологічних або біомедичних досліджень, вирішення технічних і, що дуже важливо, промислових завдань: карти контролю якості, аналіз процесів і планування експерименту.

За допомогою вбудованої мови програмування STATISTICA BASIC можна створювати рішення, які просто інтегруються до інших додатків.

Слід зауважити, що склад модулів відчутно розрізняється залежно від версії та типу ліцензії пакету. Наприклад, базова версія може додатково комплектуватися спеціалізованими модулями: Power Analysis (планування статистичних досліджень), Neural Networks (нейромереживний аналіз) і т. ін.

Перша версія пакета була створена в 1991 р. Остання версія програми 13.3 (2016 р.).

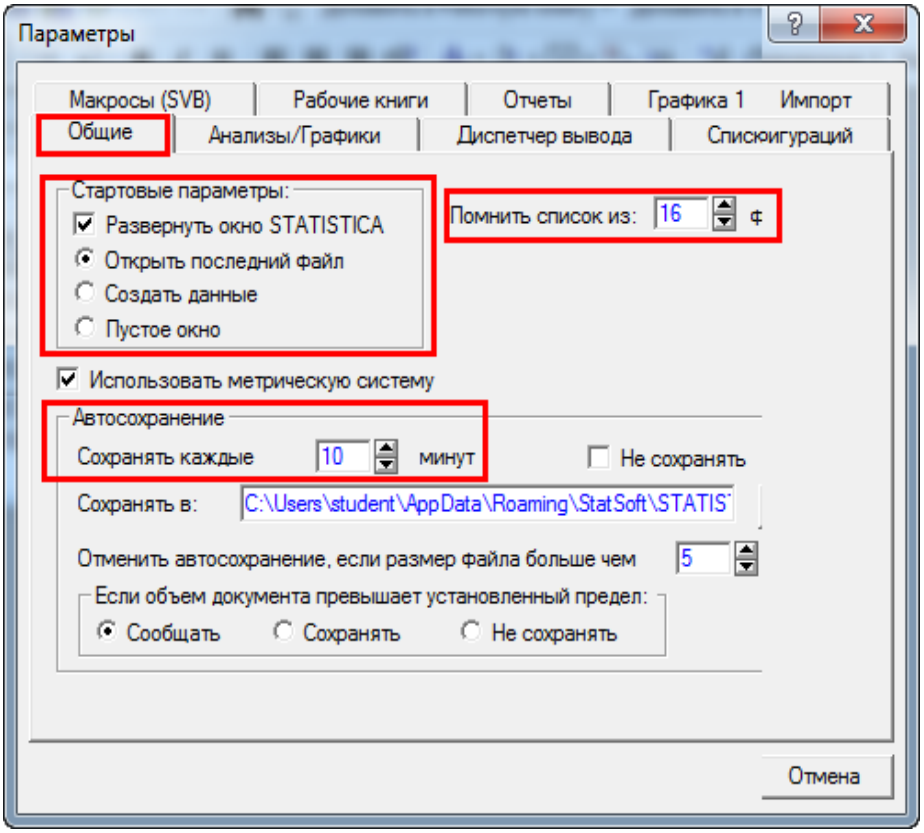
Робота із системою

У загальному випадку робота із системою передбачає таку послідовність дій.

1. Визначити структуру даних.
2. Ввести первині дані.
3. Провести дослідження даних на помилки.
4. За необхідності здійснити попереднє перетворення даних, наприклад групування або ранжування.
5. Розрахувати описові статистики.
6. Здійснити візуалізацію даних.
7. Застосувати конкретній метод аналізу.

Головне вікно програми

Початковий варіант роботи із системою визначається у настроюваннях системи, доступ до яких здійснюється у групі "Стартовые параметры" на вкладці "*Общие*" ("Загальні") вікна "*Параметры*". Це може бути створення нового файлу даних, завантаження існуючого файла і т. ін. Наприклад якщо як початковий варіант роботи у настроюваннях системи є відкриття файлу даних системи, то головне вікно міститиме останній документ, з яким відбувалося робота (звичайно, за умови якщо він доступний).



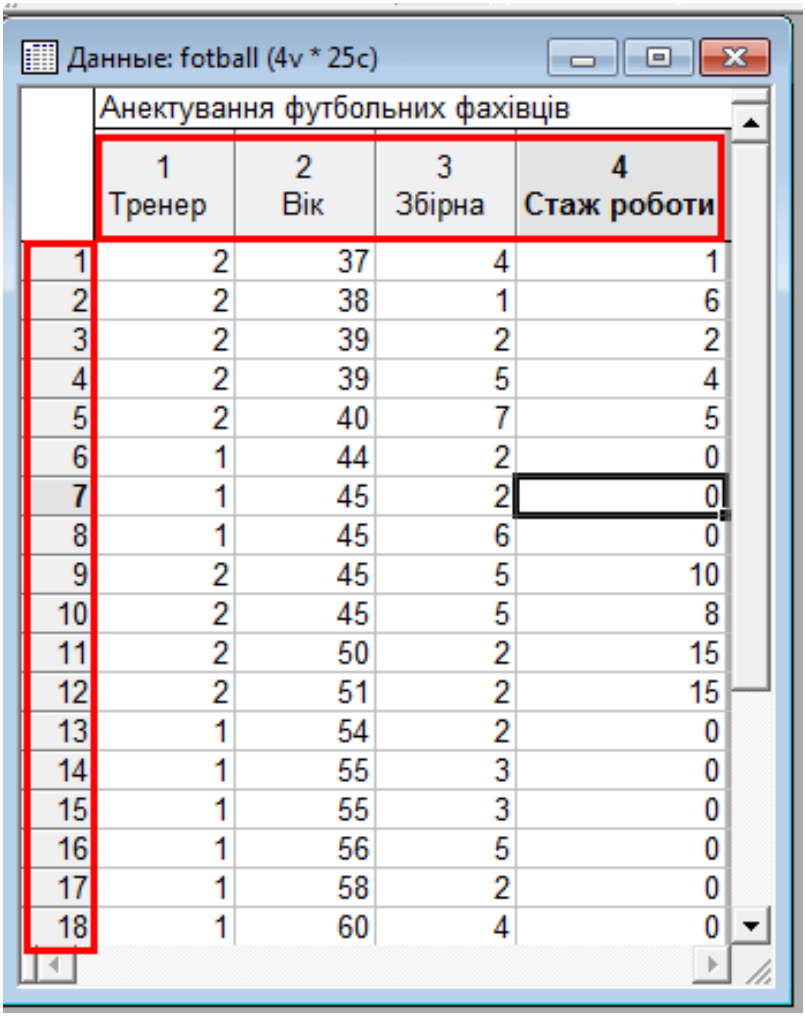
Основну область вікна займає область для введення даних, що за принципом побудови дуже схоже з документом електронної таблиці.

Введення первинних даних

Таблиця даних

Подання даних у програмі має табличний вигляд і зовні дуже схоже з електронною таблицею Excel. При цьому у рядках таблиці розташовуються *спостереження* (*Cases*), а у стовпчиках – *змінні* (*Variables*). Заголовки рядків (спостереження) нумеруються

арабськими цифрами, а заголовки стовпчиків містять ім’я змінної.


Данные: fotball (4v * 25c)

	Анектування футбольних фахівців			
	1 Тренер	2 Вік	3 Збірна	4 Стаж роботи
1	2	37	4	1
2	2	38	1	6
3	2	39	2	2
4	2	39	5	4
5	2	40	7	5
6	1	44	2	0
7	1	45	2	0
8	1	45	6	0
9	2	45	5	10
10	2	45	5	8
11	2	50	2	15
12	2	51	2	15
13	1	54	2	0
14	1	55	3	0
15	1	55	3	0
16	1	56	5	0
17	1	58	2	0
18	1	60	4	0

Наприклад, якщо до таблиці потрібно занести дані анкети, то окреме питання анкети розглядається як змінна, а окрема анкета, що містить усі питання, буде спостереженням.

У загальному випадку для створення нової таблиці з даними слід виконати команду **Файл ▸ Создать (File ▸ New)** або натиснути відповідну кнопку на панелі стандартних інструментів. З'явиться вікно "*Создать новый документ*" ("Create New Document", створення нового документа). За замовчуванням новий файл даних створюється для 10 спостережень і 10 змінних, але ці значення можна у вікні змінити.

Файли даних у системі називаються *Таблица данных* (*Spreadsheet*).

Приклад

Процес підготовки даних і їх дослідження розглядатимемо на умовному прикладі анкетування 25 футбольних фахівців на тему "Яка збірна, на Ваш погляд, стане чемпіоном на майбутньому чемпіонаті світу?" До анкети було включено такі запитання:

Таблиця 1

Структура даних анкети

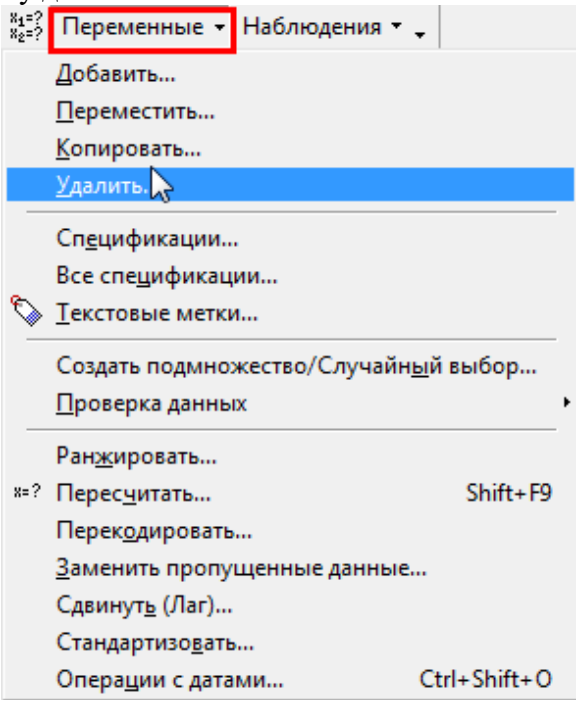
Питання анкети	**Ім'я змінної**	**Значення змінних**
Чи є Ви діючим тренером ("Так", "Ні", "відповідь відсутня")	Тренер	"2" – так, "1" – ні, "0" – дані відсутні
Вік (повних років)	Вік	
Збірна, яка стане чемпіоном на майбутньому чемпіонаті світу.	Збірна	Збірна: Аргентина – "1", Бразилія – "2", Германія, Іспанія, Італія, Франція, інші.
Стаж роботи тренером (повних років)	Стаж роботи	

Створення структури даних таблиці

1. Задати потрібну кількість спостережень і змінних для таблиці.

- Для зміни кількості змінних на панелі інструментів слід

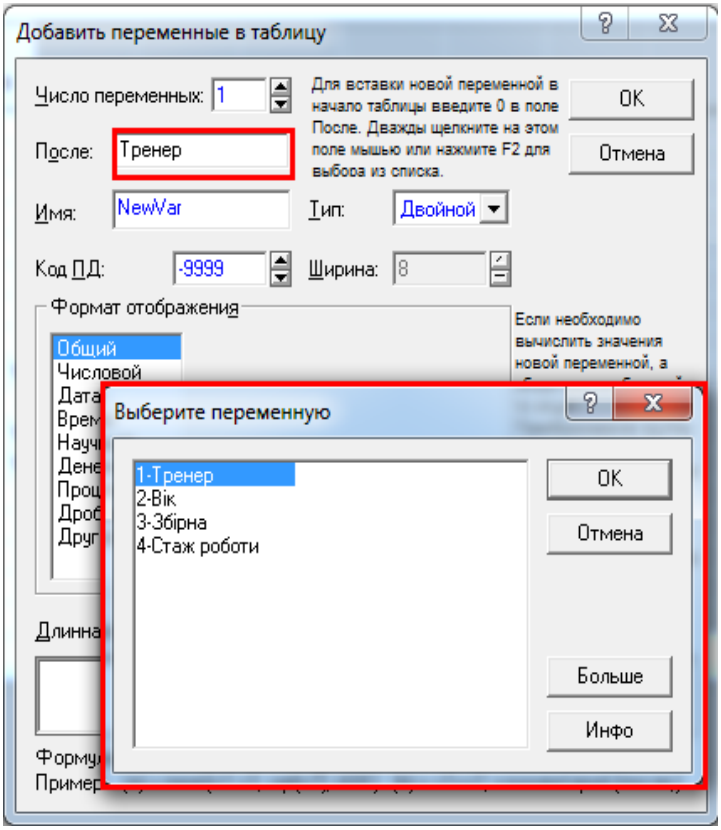
розкрити список-кнопку **«Переменные» («Vars»)** і вибрати з її меню потрібну дію.



Якщо потрібно вилучити зайві змінні, то вибирається пункт "Удалить" ("Delete"), а якщо додати – "Добавить" ("Add"). У нашому прикладі 5 змінних, тому слід вилучити 5 зайвих. Після вибору пункту "Удалить" з'явиться вікно "*Удалить переменные*" ("Delete Variables"), в якому в полях "С переменной" ("From variable") і "По переменную" ("To variable") слід вказати початкове і кінцеве значення імен змінних, що вилучаються і натиснути кнопку **«ОК»**.

 При додаванні нової змінної слід звернути увагу на ім'я змінної в полі "После" ("After", після). Якщо у цьому полі відображається ім'я не тієї змінної, після якої потрібно додати нову змінну, його слід змінити ручним введенням нового імені з клавіатури. Але значно простіше це можна зробити, викликавши список змінних

для чого слід здійснити подвійне клацання мишею в полі "После" або натиснути клавішу **<F2>**.



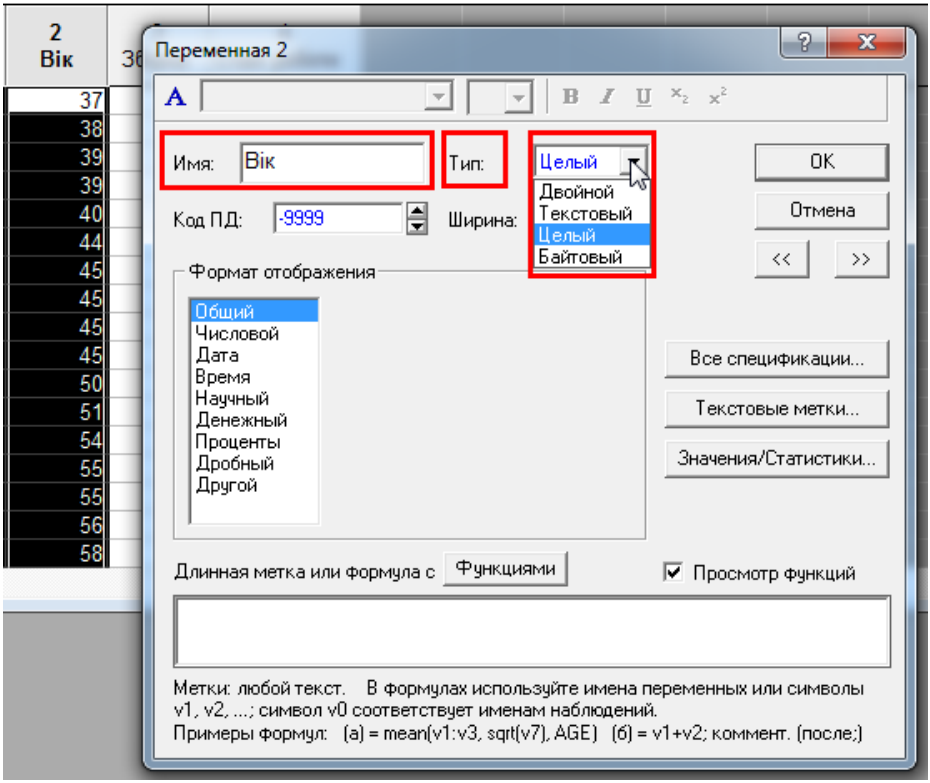
- Для зміни кількості спостережень на панелі інструментів слід розкрити список-кнопку **«Наблюдения»** (**«Cases»**, спостереження) і вибрати з її меню потрібну дію. У нашому прикладі 25 спостережень, тому слід додати 15 спостережень. Після вибору пункту "Добавить…" ("Add…") з'явиться вікно "*Добавить наблюдения*" ("Add Cases"), в якому в полях "Число наблюдений" ("Haw many") і "Вставить после" ("Insert after case") слід вказати кількість спостережень, що слід додати і після якого саме спостереження слід здійснити вставлення

нових, і натиснути кнопку **«ОК»**.

- Змінні можна додавати, вилучати, переміщувати і копіювати за допомогою контекстного меню, яке потрібно викликати на *імені змінної* у рядку заголовка. З меню вибирається потрібний елемент, наприклад "Добавить переменные", "Удалить переменные", "Переместить переменные", "Копировать переменные" (Add Variables, Delete Variables, Move Variables, Copy Variables).

2. Для кожної таблиці даних можна ввести загальну додаткову інформацію, що зазвичай використовується для ідентифікації звітів. Для цього слід здійснити подвійне натискання на полі, що знаходиться під заголовком вікна ("*Данные*:…", "Data:…").

3. Визначення властивостей змінних. Для завдання імені та інших властивостей змінної потрібно двічі натиснути на її імені у рядку заголовка. З'явиться вікно "*Переменная*" ("Variable", змінна), в якому і задаються властивості змінної.



- Поле "Имя" ("Name", ім’я змінної). За замовчуванням змінні мають імена Var… (скорочення від англ. variables – змінні).
- Поле "Тип" ("Type", тип даних). У системі визначені такі типи: "Двойной" (Double, числовий), "Целый" (Integer, цілий числовий), "Байтовый" (Byte, цілий числовий), "Текстовый" (Text, текстовий).

Числові типи розрізняються в першу чергу допустимим діапазоном значень. Наприклад, тип даних "Байтовый" (Byte) використовується для змінних, що визначаються цілими числами в діапазоні від 0 до 255 включно. Наприклад, для нашого прикладу змінна "Чи є Ви діючим тренером" може мати тільки два значення "Так" і "Ні". Закодуємо їх так: "2" – "Так", "1" – "Ні". Оскільки значення є числом з одного знака, то можна вибрати саме тип "Байтовый" або

"Целый". За необхідністю тип даних у таблиці можна відобразити у заголовку поруч з іменем змінної. Для цього слід виконати команду **Вид ▸ Имена переменных ▸ Показать типы (View ▸ Variables Headers ▸ Display Types)**. (Вигляд – імена змінних – відобразити типи даних).

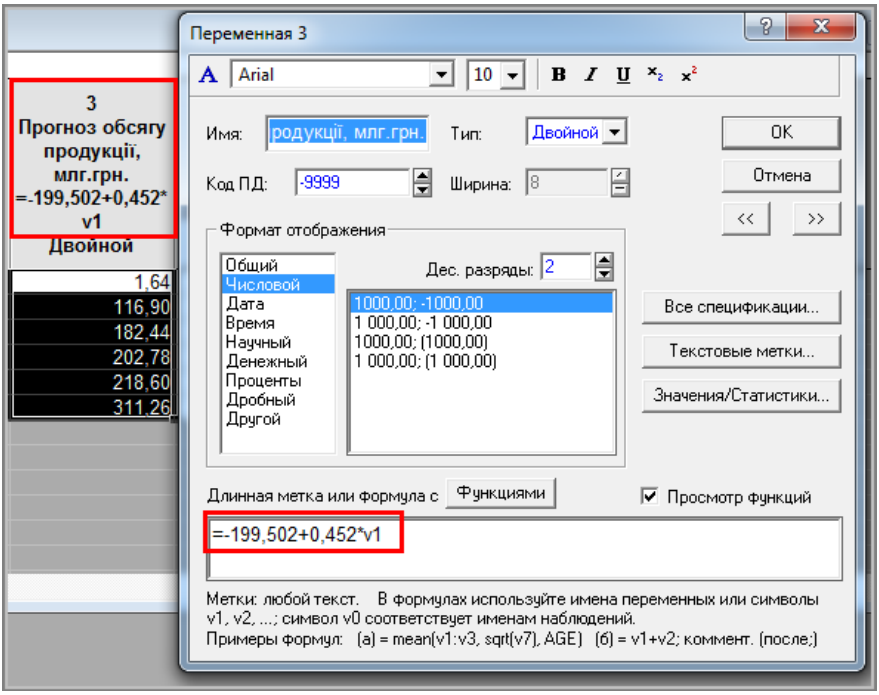
- Деталізація типу даних здійснюється у полі "Формат отображения" ("Display format", Формат відображення). Наприклад, для числового формату додатково можна встановити кількість знаків після коми, формат "Денежный" ("Currency") дає змогу відобразити дані у вигляді грошових сум і т. ін.

-  Деякі змінні можна обчислити на основі значень інших змінних. Наприклад, площа прямокутника – це добуток його ширини та висоти. Для цього застосовують розрахункову формулу. Для цього слід натиснути мишею в полі, що знаходиться під текстом "Длинная метка или формула с Функциями" ("Long name (label or formula with)") і ввести формулу. При цьому слід врахувати таке.

3.1. Формула повинна починатися зі знаку "=".

3.2. Вона може містити імена змінних, математичні і логічні операції, функції тощо. Для визначення та підстановки до формули імені вбудованої функції слід натиснути кнопку **«Функциями»**.

3.3.  Незважаючи на те, яке ім’я змінній надав користувач, воно завжди визначається літерою "v" і її порядковим номером: v1, v2, v3… Наприклад, якщо змінна v2 є сумою змінних v1 і v4, то формула буде мати вигляд =v1+v4.

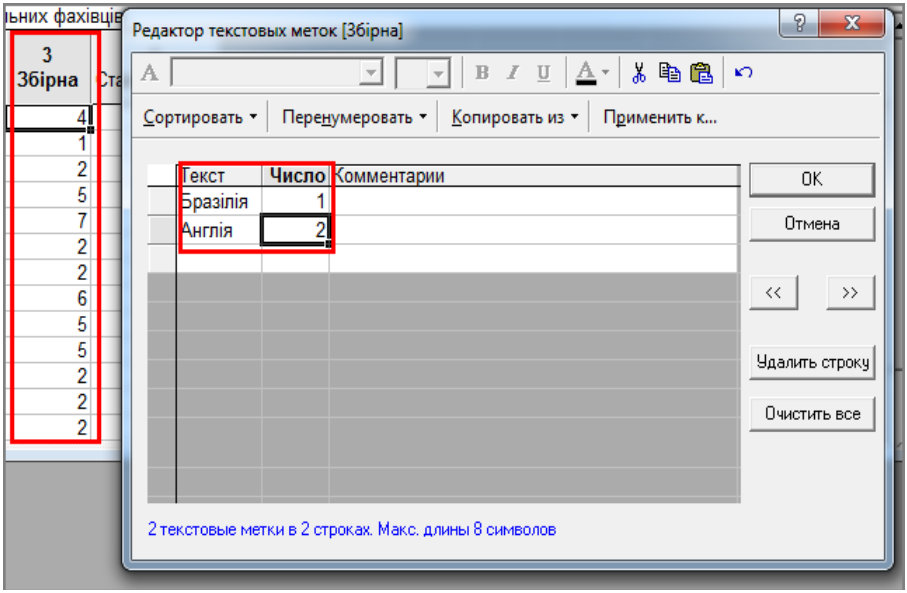


3.4. За необхідністю розрахункову формулу можна відобразити у заголовку. Для цього слід виконати команду **Вид ▸ Имена переменных ▸ Показать длинные имена (View ▸ Variables Headers ▸ Display Longs Names)**. (Вигляд – імена змінних – відображати довгі імена).

- Для того, щоб можна було виконувати статистичну обробку якісних (не числових) змінних їм надають числовий тип, і здійснюють їх кодування числами. Для кращого розуміння таких даних доцільно поставити у відповідність їхнім числовим еквівалентам текстові значення. Це здійснюється так.

3.1. У вікні з властивостями змінної слід натиснути кнопку **«Текстовые метки»** ("Text Labels", текстові мітки). З'явиться вікно "*Редактор текстовых меток*" ("Text Labels Editor").

3.2. У стовпці "Текст" ("Text Labels") ввести текстове значення змінної, а у стовбці "Число" ("Numeric") – її числовий еквівалент.



Надалі це дозволяє перемикати відображення даних у таблиці з числових значень на текстові і навпаки за допомогою кнопки панелі інструментів  "Показать/Скрыть текстовые метки" ("Display Text Label") або за командою **View ▸ Display Text Label (Вид ▸ Показать текстовые метки)**.

Введення первинних даних

При введенні даних слід врахувати таке:

1. Клітинки з числовим типом обов'язково повинні містити числове значення, навіть якщо це значення "0". При невиконанні цієї умови буде з'являтися повідомлення "Переменная не

содержит дисперсию" ("Змінна не містить дисперсію").

2.  Не потрібно вводити дані для змінних, що можна обчислити за розрахунковою формулою на основі інших змінних. При цьому значення у формулах після змін даних може бути розраховано автоматично або в ручному режимі. В ручному режимі для цього слід виконати команду **Данные ▸ Пересчитать** або натиснути клавішу **<F9>**. Різниця між цими діями в тому, що за першим варіантом користувач одержує можливість встановлення перерахунку для певних клітинок, а за другим здійснюється автоматичний перерахунок усіх даних. Для автоматичного розрахунку даних за формулою при редагуванні даних слід здійснити настроювання системи:

- Виконати команду **Сервис ▸ Параметры (Tools ▸ Options)**. Відкриється вікно "*Параметры*" ("Options").
- Перейти на вкладку "*Таблицы*" ("Spreadsheet").
- Встановити позначку для поля-мітки "Автоматически пересчитывать формулы при изменении данных" ("Auto-recalculate spreadsheet formulas when data change", "Виконувати автоматичний розрахунок даних за формулами при зміні даних").

3. Якщо текстовим даним поставити у відповідність числові еквіваленти, тобто створити текстові мітки, то дані можна вводити у текстовому форматі, а система надалі самостійно перетворить ці дані в їх числові еквіваленти.  Можна взагалі попередньо не задавати для текстових значень числові еквіваленти, оскільки введення до клітинки текстового значення автоматично створює для неї числовий еквівалент.

Сортування спостережень

Спостереження у таблиці можна впорядковувати (сортувати) за значеннями однієї або відразу кількох змінних. Для цього слід виконати такі дії.

1. Виконати команду **Данные ▸ Сортировка** (**Data ▸ Sort**). Відкриється вікно "*Параметры сортировки*" ("*Sort Options*").

2. Визначити змінну (або змінні), за якою слід впорядкувати дані, напрямок сортування (за зростанням або за спаданням). Для змінних текстового типу можна також вибрати варіант сортування: за текстом або за числовим кодом, що відповідає текстовому опису змінної.

3. Натиснути **«ОК»**.

Сервісні функції

Система дає змогу швидко відобразити усі властивості однієї або всіх змінних. Для цього слід встановити курсор на стовпчик, що містить потрібну змінну і виконати команду **Данные ▸ Спецификации переменной** або **Данные ▸ Все спецификации переменных (Data ▸ All Variables Spec**).

За необхідності зміни розміру клітинки, наприклад якщо значення не поміщається в стовпчику, це здійснюється засобами, зазвичай прийнятими в електронних таблицях, тобто перетягуванням межі стовпчика або подвійним натисканням на правій межі його заголовка.

Якщо дані були сформовані в інших статистичних пакетах, наприклад, SPSS або в електронній таблиці, то їх можна скопіювати до таблиці з даними за допомогою стандартної дії вставки. Наприклад, виділити у *SPSS* потрібні дані, скопіювати їх до буферу обміну, а потім у *STATISTICA* виконати команду **Правка ▸ Вставить (Edit ▸ Paste)** або натиснути  на стандартній панелі інструментів. Зрозуміло, що за таким варіантом для змінних з нечисловим типом даних зберігаються тільки числові значення і не зберігаються їхні текстові значення (наприклад, для значення "Тренер" запам’ятається "1").

Збереження даних

Збереження файлів відбувається стандартним чином за командою **Файл ▸ Сохранить (File ▸ Save)** або натисканням піктограми  . Якщо потрібно зберегти існуючий файл з іншим ім’ям, то це

здійснюється за стандартною командою **Файл ▸ Сохранить как… (File ▸ Save As…)**. До імені файлу *STATISTICA* автоматично додає розширення *STA*. Такий файл даних є набором файлів, оскільки він містить і автоматично зберігає інформацію про всі додаткові файли (графіки, звіти, програми), що використовуються з поточним набором даних. Крім цього зберегти таблицю даних можна також у форматі електронної таблиці *Excel*, у вигляді веб-сторінки, текстового файлу і т. ін.

Відкриття даних

За певних настроювань під час завантаження системи автоматично відкривається останній файл, з яким відбувалося робота. Відкрити файл даних можна й під час роботи з системою за командою **Файл ▸ Открыть (File ▸ Open ▸ Data)**.

При цьому в робочій області може бути тільки один файл з даними.

Під час роботи з даними доцільно встановити режим автоматичного збереження інформації. Для цього слід виконати команду **Сервис ▸ Параметры (Tools ▸ Options)** і на вкладці "*Общие*" ("Загальні") вікна "*Параметры*" встановити числове значення для поля "Сохранять каждые ... минут" ("Зберігати кожні ... хвилин").

Дослідження даних

На першому кроці дослідження доцільно піддати докладному аналізу самі дані з метою виявлення помилок введення, а також здійснити перевірку закону розподілу даних на нормальність.

1. Виявлення помилок введення

Найточніший спосіб перевірки даних (тобто значень всіх змінних) на помилки полягає у звірці даних, уведених у таблицю, з оригіналом (наприклад, з анкетою). Проте цей спосіб вимагає дуже багато часу, особливо за великого обсягу даних. Тому

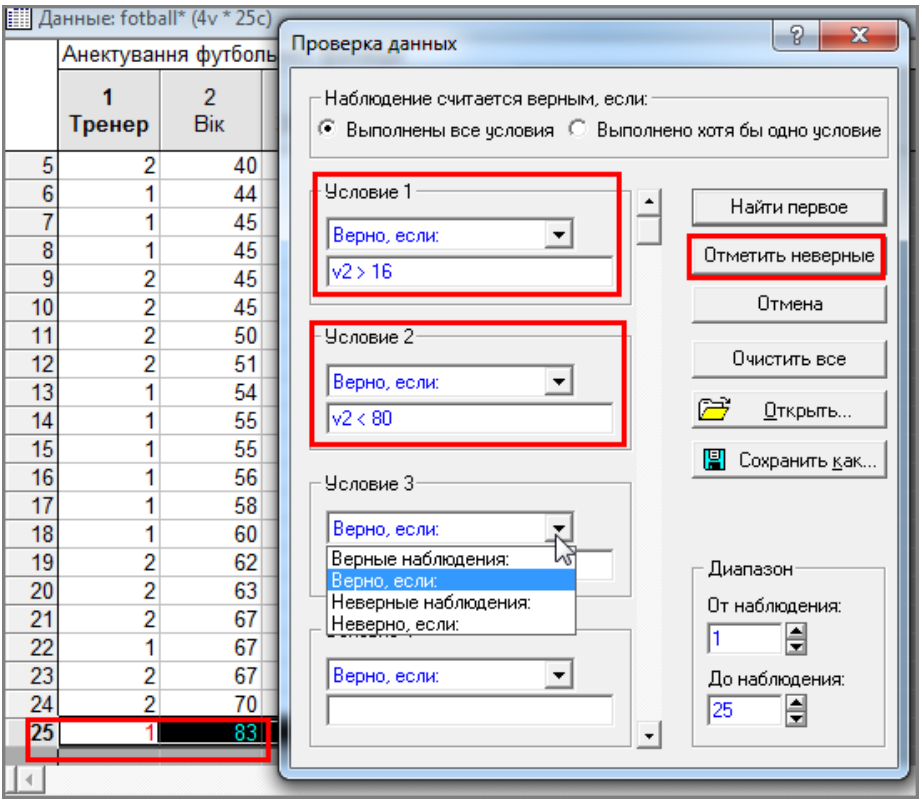
проводити таку трудомістку роботу доцільно тільки коли обсяг даних є невеликим. У загальному випадку рекомендується проводити частотний аналіз значень змінних. Результати частотного аналізу досить часто дозволяють виявити невірні значення. Наприклад, якщо змінна містить зріст в сантиметрах, то значення "400" при частотному аналізі явно свідчить про те, що в даних є помилка. Після проведення частотного аналізу це значення можна відшукати у таблиці даних і виправити. Отже, під час аналізу частотних таблиць особливу увагу треба звертати на максимальні і мінімальні значення. Проте, якщо замість віку 65 років було введено, наприклад, значення 56, то за допомогою таблиці частот цю помилку виявити неможливо.

Часто є можливість провести смисловий аналіз даних за допомогою таблиць спряженості. Наприклад, якщо дані узяті з анкети, в якій було питання щодо сімейного стану (холостий/не заміжня, одружений/одружена, вдівець/вдова, розведений(а)), то, побудувавши таблицю спряженості для цього питання і питання на зразок: "Якщо у вас є сім'я, то чи прийнятно у вас проводити відпустку окремо один від одного?", легко можна виявити, чи відповіли на нього тільки одружені опитувані.

Маючи певні практичні навички, за допомогою таких прийомів можна виявити велику кількість помилок введення. Всі такі помилки обов'язково повинні бути виправлені. Навіть якщо спостережень кілька тисяч, то навіть одне суперечливе значення завдає шкоди дослідженню, тому що створюється враження, що робота із збирання інформації виконана неякісно.

Дослідження даних може бути проведено і спеціальними засобами програми. Для виявлення помилок введення програма має спеціальний модуль "Перевірка даних", доступ до якого здійснюється за командою **Данные ▸ Проверка данных (Data ▸ Verify Data)**. У вікні модуля можна задати для даних кілька умов, виконання яких (усіх або хоч би одного) дозволяє вважати їх правильними. Кожна умова складається з двох частин. Перша частина – це список, що містить заздалегідь визначені правила,

наприклад, "Верно, если", "Верные наблюдения". Друга частина є виразом, як правило, це умова. Умова формується за такими правилами.



1. Ліва частина умови – це ім’я змінної, яке позначається літерою "v" із додаванням її порядкового номера в таблиці даних.

2. Після імені змінної вводиться логічний оператор.

3. Права частина умови – це значення з яким порівнюється змінна. При цьому значенням може бути число, математичний вираз, функція тощо.

Наприклад, значення змінної "Вік" для тренера з високою ймовірністю буде в межах від 16 до 80 років. Якщо у таблиці з

даними ця змінна має ім’я "v2", то це припущення можна виразити через дві умови :

1. v2 > 16
2. v2 < 280

Після формування умов також потрібно визначити таке:

1. *Діапазон перевірки*. За замовчуванням перевіряються всі спостереження, але можна задати перевірку на певному інтервалі спостережень.

2. *Характер перевірки*. Знайти відразу всі помилкові значення або робити це крок за кроком, відшукуючи спостереження з неправильними даними послідовно одне за одним. Це здійснюється відповідно за допомогою інструментів "Отметить неверные" і "Найти первое". За першим варіантом усі неправильні значення виділяються червоним кольором.

2. Перевірка закону розподілу на нормальність

Щодо нормальності розподілу можна судити вже візуально з графіку у вигляді гістограми. Якщо побудувати графік за емпіричними даними і за очікуваними нормальним розподілом, то візуально достатньо просто визначити чи є розподіл нормальним. Чим ближче один до одного розташовані графіки, тим ближче характер емпіричного розподілу до нормального.

Але строге математичне підтвердження нормальності розподілу здійснюється з використанням спеціальних статистичних критеріїв нормальності. У *STATISTICA* з цією метою застосовують критерії Колмогорова-Смірнова/Лілієфорса і Шапіро-Уїлка.

Критерій Колмогорова-Смірнова (позначається D) базується на максимальній відмінності між емпіричною функцією і теоретичною функцією розподілу. Якщо статистика D є значущою, то гіпотеза, що емпіричний розподіл є нормальним, має бути відхилена. Для багатьох програм, розраховані значення ймовірності мають силу, коли середнє та стандартизоване відхилення нормального розподілу відоме апріорі та не оцінене за даними. Проте зазвичай ці характеристики обчислені за фактичними даними. У

цьому випадку критерій нормальності містить складну умовну гіпотезу і використовується ймовірність Лілієфорса (Lilliefors probabilities). Проте останнім часом для перевірки нормальності розподілу прийнято використовувати критерій Шапіро-Уїлка, а критерій Колмогорова-Смірнова реалізований у STATISTICA через його історичну популярність.

Критерій Шапіро-Уїлка (позначається як *W*) зараз є привілейованим критерієм на нормальність, оскільки його властивості у більшості випадків мають помітно більшу потужність над альтернативними критеріями.

Статистика W обчислюється за формулою:

$$W = \frac{b^2}{S^2}, \tag{1}$$

де

$$S^2 = (x - \tilde{x})^2 \tag{2}$$

$$b = \Sigma a_{n-i+1}(x_{n-i+1} - x_i) \tag{3}$$

де $\tilde{x}$ – середня вибірки;

a_{n-i+1} – табличні константи.

У *STATISTICA* при виконанні перевірки розподілу на нормальність за критерієм Шапіро-Уїлка розраховуються дві статистики: значення коефіцієнта *W* і рівень значущості *p*.

Сам тест на нормальність за критерієм Шапіро-Уїлка є перевіркою нульової гіпотези про те, що емпіричний розподіл не відрізняється від очікуваного теоретичного нормального розподілу. При цьому значення *W* прагне наблизиться до "1" для будь-яких рівнів значущості *p*. Чим ближче значення *W* наближається до "1", тим менше вірогідність помилково прийняти гіпотезу щодо нормальності розподілу. Альтернативною гіпотезою є те, що розподіл відрізняється від нормального. При цьому значення *W* прагне наблизиться до "0" при рівні значущості $p < 0{,}05$.

Під час використання критерію потрібно звертати увагу не тільки на значення коефіцієнта *W*, але й на статистичний рівень

значущості p. Оскільки нульова гіпотеза сформульоване про те, що розподіл є нормальним, то вона буде прийматися, коли рівень статистичної значущості $p > 0{,}05$, а значення коефіцієнта $W > 0{,}9$, тобто він має високе значення. У протилежному випадку приймається альтернативна гіпотеза.

Слід мати на увазі, що у *STATISTICA* використовується вдосконалена версія алгоритму розрахунку критерію (Royston, 1982 р.), яку можна застосовувати до вибірок, що містять від 8 до 2000 спостережень (традиційно вважається, що застосування критерію Шапіро-Уїлка обмежується вибірками до 50 спостережень).

Перевірити розподіл на нормальність можна різними шляхами. Наприклад, дуже просто це можна зробити в модулі "Основные статистики и таблицы" ("Bases Statistics/Tables") під час побудови гістограми:

1. У вікні модуля вибрати пункт "*Описательные статистики*" ("Descriptive Statistics:").

2. У вікні описових статистик перейти на вкладку "Нормальность" ("Normality").

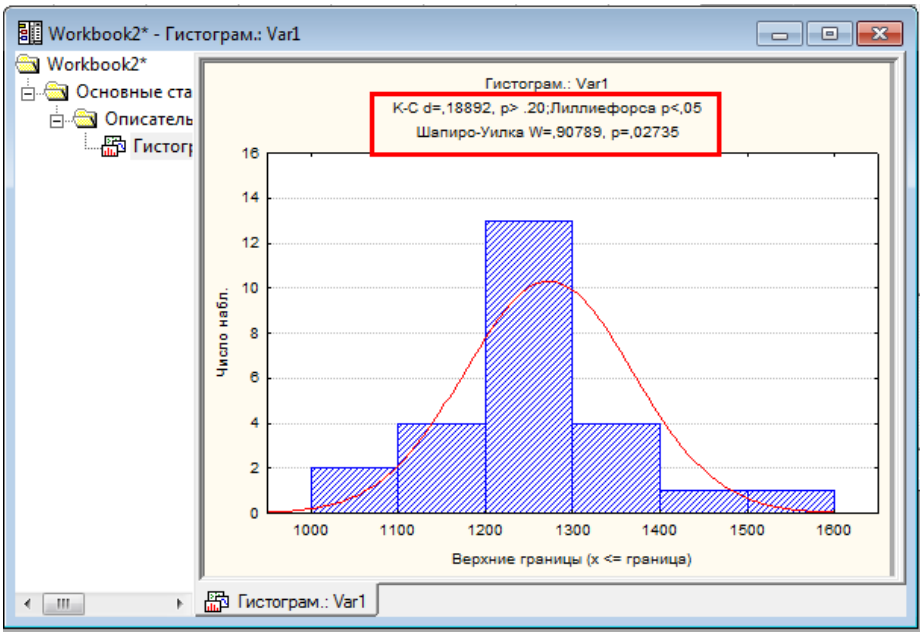
3. Якщо змінна є неперервною, то в полі "Число интервалов" ("Number of intervals") можна змінити кількість інтервалів групування.

4. Для відображення очікуваних теоретичних частот разом з фактичним розподілом слід встановити прапорець для поля-мітки "Ожидаемые нормальные частоты" ("Normal expected frequencies").

5. Для розрахунку потрібного критерія на нормальність встановити прапорець біля назв критеріїв "Критерий нормальности Колмогорова-Смирнова и Лиллиефорса" ("Kolmogorov-Smirnov & Lilliefors test for normality") та (або) "Критерий Шапиро-Уилка" ("Shapiro-Wilk's W test").

6. Натиснути кнопку **«Гистограммы»** (**«Histograms»**).

Розраховані критерії відображаються у верхній частині вікна з гістограмою.



Розрахунок базових статистичних показників

Вибір модуля (метода статистичного аналізу)

Як зазначалося раніше, *STATISTICA* містить кілька модулів, кожний з яких реалізує певний метод статистичного аналізу. Звернутися до цих модулів можна кількома способами:

1. використовуючи пункти головного меню **Анализ (Statistics)**;
2. використовуючи панель інструментів "Анализ" ("Statistics"). Якщо вона відсутня, то доцільно вивести її на екран. Це здійснюється стандартним чином: за командою **Вид ▸ Панели инструментов ▸ Анализ (View ▸ Toolbar ▸ Statistics)** або викликом контекстного меню в області панелі інструментів і вибором з меню на пункту "Анализ" ("Statistics");
3. натиснути кнопку виклику меню для найбільш вживаних інструментів, що знаходиться у лівому куті панелі статусу 

("Вызвать меню часто используемых средств", "Start menu").

Робочі книги

Всі результати, у тому числі графіки, записуються у *робочі книги* (*Workbook*).  При цьому результати будь-якого аналізу відкритої таблиці даних записуються в одну робочу книгу, але на окремому аркуші. Список одержаних результатів відображається в лівій частині вікна.

Самі робочі книги можна зберігати та редагувати, вставляти, вилучати, перейменовувати окремі аркуші. Для виконання певної дії слід викликати контекстне меню на назві аркуша лівої частини вікна і вибрати з нього потрібну дію.

Окремі аркуші книги можна "витягувати" з неї, після чого вони вилучаються з неї і з'являються у новому вікні. Для цього їх слід просто перетягнути мишею зі списку робочої книги за межі її вікна.

Частотний аналіз

Для побудови групування слід виконати такі дії:

1. Виділити змінну або змінні, для якої слід здійснити групування. Виділення рядків або стовпчиків у таблиці з даними здійснюється так саме, як і в електронних таблицях. Наприклад, виділити повністю рядок або стовпчик можна натиснувши на її заголовку, для виділення кількох несуміжних змінних натискають на їх заголовку, утримуючи натиснутою клавішу **<Ctrl>** і т.д. Проте можна і не виділяти повністю змінну: достатньо просто встановити курсор у клітинці з потрібною змінною або виділити у рядку клітинки зі змінними, якщо показники потрібно розрахувати відразу для кількох змінних.

2. Завантажити модуль "Основные статистики и таблицы" ("Bases Statistics/Tables"). Це може бути здійснено різними варіантами:

- Виконати команду **Анализ ▸ Основные статистики и таблицы (Statistics ▸ Bases Statistics/Tables)**.
- На панелі "Анализ" ("Statistics") натиснути кнопку  "Основные статистики и таблицы" ("Bases Statistics/Tables").

• Натиснути кнопку  "Вызвать меню часто используемых средств" ("Start menu", "Викликати меню засобів, що найчастіше використовуються") і вибрати з меню **Анализ ▸ Основные статистики и таблицы (Statistics ▸ Bases Statistics/Tables)**.

З'явиться вікно "*Основные статистики и таблицы*" ("Basic Statistics and Tables:").

3. Вбрати у вікні пункт "Таблицы частот" ("Frequency tables") і натиснути **«ОК»**. З'явиться вікно "*Таблицы частот*".

4. Групування даних здійснюється за параметрами, які знаходяться у вікні модуля у групі "Метод категоризации для таблиц и графиков" (Categorization method for tables & graphs) на вкладці "*Дополнительно*" ("Advanced"). За замовчуванням створюється окрема група для кожного значення змінної, що визначається встановленим значенням перемикача в положення "Все значения отдельно" (All distinct value). Але можна застосувати і інші параметри побудови групування:

• Для дискретних числових ознак або описових ознак, що також задаються дискретними числовими значеннями, слід встановити перемикач для поля-мітки "Группировка" в положення "Целые интервалы (категории)" (Integer Categories).

• Для неперервних ознак побудову групування можна здійснити за інтервалами, використовуючи такі параметри:

4.1. Задати кількість (рівних) інтервалів в полі "Число равных интервалов" (No. of exact intervals).

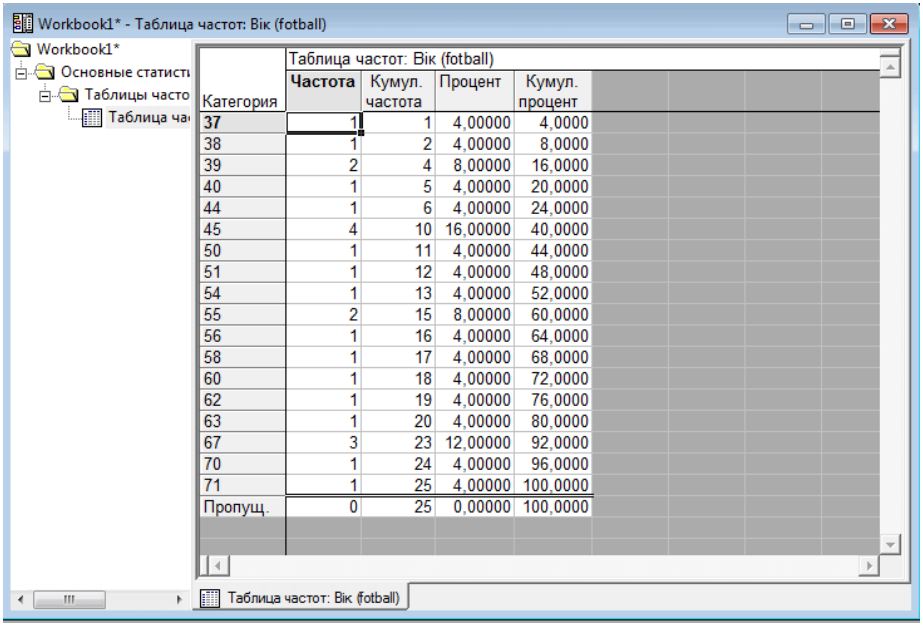
4.2. Побудувати групування, де межі інтервалів є кратними "10", для чого використовується параметр "Приблизительное число интервалов" ("Near" intervals; approximate no.).  При цьому слід мати на увазі, що за таким способом фактична кількість інтервалів може не співпадати із заданою кількістю.

4.3. Групування можна здійснити шляхом визначення ширини інтервалу, встановивши перемикач в положення "Размер шага" (Step size) і задати в полі праворуч від

нього початкове значення першого інтервалу. Воно може бути автоматично визначене як мінімальне з усіх значень або нульовим. Для застосування першого варіанту достатньо встановити прапорець для поля-мітки "с мин. знач." (at minimum, з мінімального значення).

5. Модуль завжди розраховує групові і кумулятивні частоти, проценті тощо. Перелік інших показників, що розраховуються та виводяться у робочу книгу, визначається на вкладці "*Опции*" (Options).

6. На вкладці "*Быстрый" ("Quick"*) натиснути кнопку **«Таблицы частот»** (**«Frequency Tables»**) або **«Гистограммы»** (**«Histograms»**) відповідно для табличної або графічної побудови ряду розподілу.



Категория	Частота	Кумул. частота	Процент	Кумул. процент
37	1	1	4,00000	4,0000
38	1	2	4,00000	8,0000
39	2	4	8,00000	16,0000
40	1	5	4,00000	20,0000
44	1	6	4,00000	24,0000
45	4	10	16,00000	40,0000
50	1	11	4,00000	44,0000
51	1	12	4,00000	48,0000
54	1	13	4,00000	52,0000
55	2	15	8,00000	60,0000
56	1	16	4,00000	64,0000
58	1	17	4,00000	68,0000
60	1	18	4,00000	72,0000
62	1	19	4,00000	76,0000
63	1	20	4,00000	80,0000
67	3	23	12,00000	92,0000
70	1	24	4,00000	96,0000
71	1	25	4,00000	100,0000
Пропущ.	0	25	0,00000	100,0000

Описові статистики

За допомогою *описових статистик* визначаються найбільш загальні властивості емпіричних даних, які дають загальне уявлення відносно значень, що набуває змінна. До них належать се-

редня, вибіркова дисперсія, стандартне відхилення, медіана, мода, максимальне та мінімальне значення, розмах варіації та квартилі.

Для розрахунку описових статистик слід виконати такі дії.

1. Виділити змінну або змінні, для якої слід розрахувати показники. Для цього достатньо просто клацнути на заголовку змінної. Але вибір змінної можна здійснити і пізніше у вікні модуля, що містить кнопку **«Переменные» («Variables»)**. Натискання цієї кнопки ініціює появу вікна "*Выберите переменные для анализа*" ("Select the variables for the analysis"). Після вибору змінної праворуч від кнопки **«Переменные»** відображається ім'я вибраної змінної. Надалі використовуйте цей алгоритм для вибору змінної або змінних.

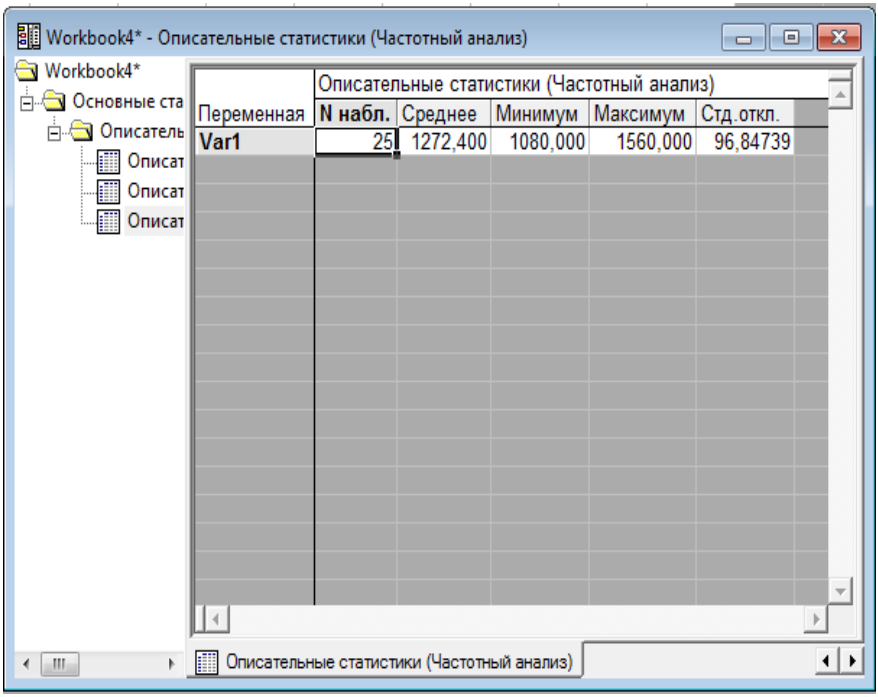
2. Завантажити модуль "Основные статистики и таблицы" ("Bases Statistics/Tables"). З'явиться вікно "*Основные статистики и таблицы*" ("Basic Statistics and Tables:").

3. Вибрати зі списку вікна пункт "Описательные статистики" ("Descriptive Statistics", описові статистики).

4. Натиснути **«ОК»**. З'явиться вікно "*Описательные статистики*" ("Descriptive Statistics:").

5. За замовчуванням розраховуються середня, стандартне відхилення, а також визначаються кількість спостережень, максимальне і мінімальне значення. Якщо потрібно розрахувати інші описові статистики, то слід перейти на вкладку "*Дополнительно*" ("*Advanced*") і встановити позначку біля показників, що потрібно розрахувати.

6. Розрахунок показників ініціюється натисканням кнопки **«ОК»**, після чого з'явиться вікно з таблицею результатів на ім'я "*Описательные статистики*" ("Descriptive Statistics"). Розрахунок також ініціюється натисканням кнопки **«Подробные описательные статистики» («Summary: Descriptive statistics»)**, що розташована, зокрема на вкладці *"Быстрый" ("Quick"*).



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 звертаючись при цьому до первинних даних.

При переході до іншого режиму вікно "*Описательные статистики*" ("Descriptive Statistics") не закривається, а згортається на панель стану. За необхідності до нього можна надалі в будь-який момент звернутися, навіть якщо закрити вікно з результатами. При цьому у вікні зберігаються всі встановлені раніше параметри. Якщо вікно з результатами аналізу не закривати, то спроба повторного звернення до модуля "Основные статистики и таблицы" ("Bases Statistics/Tables") призведе до появи вікна-попередження, в якому вказується, що аналіз такого типу вже виконується, і буде запропоновано або продовжити цей аналіз або завантажити новий.

За необхідністю під час розрахунку описових статистик можна виконати додаткові дії:

1. Здійснити групування даних. Інструмент (кнопка) **«Таблицы частот» («Frequency tables»)** дозволяє побудувати таблиці частот.

2. Побудувати графік за допомогою інструмента (кнопки) **«Гистограммы»** («Histograms»). При цьому на вкладці "*Нормальность*" ("Normality") і "*Диаграммы*" ("Prob. & Scatterplots") можна задати додаткові параметри побудови гістограм і діаграм розсіяння, наприклад, побудову 3-вимірної гістограми ("3М гистограммы", "3D histograms").

3. Розрахувати критерії для перевірки нормальності Колмогорова-Смірнова і Шапіро-Уїлка, а також одержати теоретичні частоти для нормального розподілу.

Описові статистики можна розрахувати для будь-якого діапазону даних, в якості якого може бути діапазон клітинок, одна або кілька змінних (стовпчиків), одне або кілька спостережень (рядків) і навіть одна клітинка таблиці. Це здійснюється за допомогою так званих *блокових статистик* (*Statistics of Block Data*) і виконується за таким алгоритмом:

1. Виділити потрібні дані.

2. Виконати команду **Анализ ▸ Блоковые статистики ▸ По столбцам (По строкам) ▸ … (Statistics ▸ Statistics of Block Data ▸ Block Columns (Block Rows) ▸ …** Якщо слід розрахувати всі показники, то слід вибрати з меню останній пункт "Все" ("All"), якщо – якісь конкретний, то слід вибрати з меню потрібний. Ті самі дії можна виконати, якщо викликати контекстне меню на виділених даних і вибрати послідовно пункти **Блоковые статистики ▸ По столбцам (По строкам) ▸ … (Statistics of Block Data ▸ Block Columns (Block Rows) ▸ …)**

Статистичні графіки

Побудова графіків, тобто *здійснення візуалізації даних* є наступним етапом проведення статистичного дослідження. Багато закономірностей важко виявити безпосередньо за даними, проте

вони чітко можуть проявлятися на графіку. Візуальні методи в STATISTICA діляться на категорізовані і не категорізовані. У категорізованих методах використовуються різні способи групування даних. Завжди пам'ятаєте золоте правило аналізу даних: групуйте дані, розбивайте їх на однорідні групи, завдяки чому закономірності стають більш очевидними.

Доступ до інструментів побудови графіків здійснюється з пункту головного меню **Графика** або з панелі "Графика" (Graphs). За відсутністю панелі вона відображається стандартним чином за командою **Вид ▸ Панели инструментов ▸ Графика (View ▸ Toolbar ▸ Graphs)** або викликом контекстного меню в області панелі інструментів і натискання в меню на пункті "Графика" ("Graphs").

Ще один варіант швидко дістатися до інструментів побудови графіків здійснюється з меню кнопки  (засоби, що часто використовуються).

Найпростішими статистичними графіками є *гістограма* і *діаграма розсіяння*.

Гістограма

Гістограма будується для *однієї* змінної.

Для побудови гістограми слід здійснити одну з дій.

1. Звернутися до модуля побудови гістограми:

- виконати команду **Графика ▸ Гистограммы (Graphs ▸ Histograms)**;
- на панелі "Графика" ("Graphs") натиснути кнопку **«2M Гистограммы»** (**«2D Histograms»**);
- натиснути кнопку засобів, що найчастіше використовуються і вибрати з меню **Графика ▸ Гистограммы (Graphs ▸ Histograms)**.

Будь-яка дія спричиніть появу вікна "*2M Гистограммы*" ("2D Histograms").

2. Натиснути кнопку **«Переменные»** (**«Variables»**) і вибрати змінну.

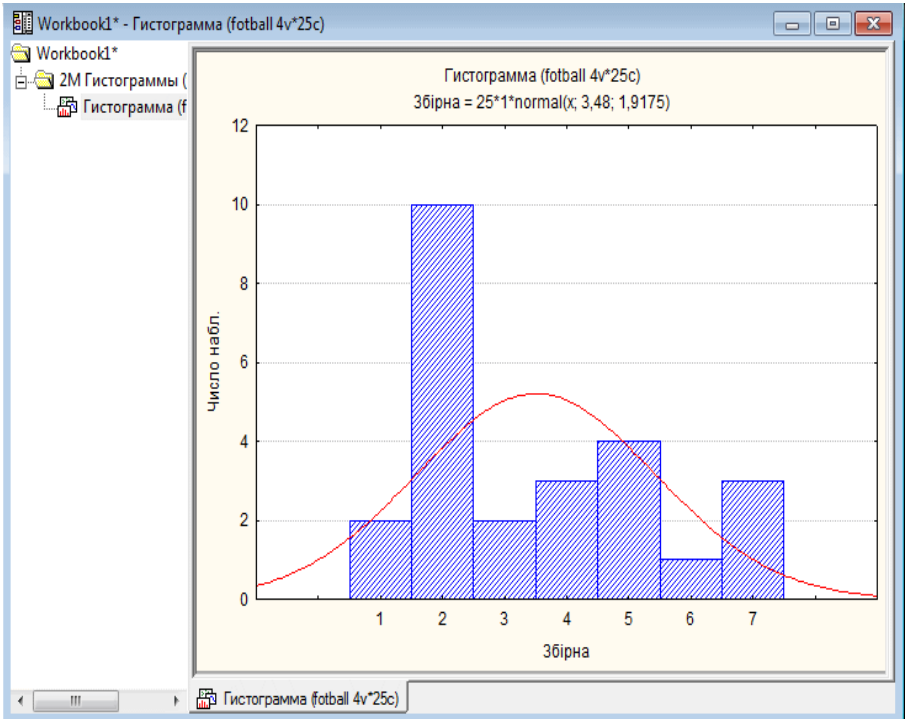
3. На вкладці "*Быстрый*" ("Quick") в полі "Категории" ("Categories") за необхідності задати потрібну кількість груп.

4. За необхідності перевірки розподілу на нормальність слід здійснити таке:

- На вкладці "*Быстрый*" ("Quick") для відображення теоретичних частот встановити позначку біля поля-мітки "Нормальное" ("Normal") у групі "Тип подгонки" ("Fit Type").
- На вкладці "*Дополнительно*" ("Advanced") для розрахунку критеріїв перевірки на нормальність Колмогорова-Смірнова і Шапіро-Уїлка в групі "Статистики" ("Statistics") встановити прапорець біля назви критеріїв.

5. Разом із гістограмою можна розрахувати і базові описові статистики. Для цього на вкладці "*Дополнительно*" ("Advanced") у групі "Статистики" ("Statistics") слід встановить позначку для поля-мітки "Описательные статистики" ("Descriptive Statistics").

6. Натиснути кнопку **«ОК»**.



Для друку графіка у його вікні слід викликати контекстне меню за його межами і вибрати з нього пункт "Печать графика" ("Print Graph") або натиснути стандартну для друку комбінацію функціональних клавіш **<Ctrl>+<P>**. За необхідності графік можна зберегти. При цьому він може бути збережений як у спеціальному форматі *STATISTICA* з розширенням STG, так і у багатьох поширених графічних форматах, зокрема ***GIF, JPEG, TIFF, BMP***.

Діаграма розсіяння

Діаграма розсіяння (діаграма кореляції, поле кореляції, scatterplot) використовується для відображення даних двох змінних, одна з яких є чинником, а інша – наслідком. За допомогою діаграми розсіяння будується графічне відображення пар даних у вигляді множини точок ("хмари") на координатній площині, що дозволяє оцінити зв'язок між двома змінними.

Алгоритм побудови графіка:

1. Здійснити одну з дій:

- Виконати команду **Графика ▸ Диаграммы рассеяния… (Graphs ▸ Scatterplots)**.
- На панелі "Графика" (Graphs) натиснути кнопку "2M Диаграммы рассеяния" ("2D Graphs").
- Натиснути кнопку засобів, що найчастіше використовуються і вибрати з меню **Графика ▸ Диаграммы рассеяния…** (**Graphs ▸ Scatterplots)**.

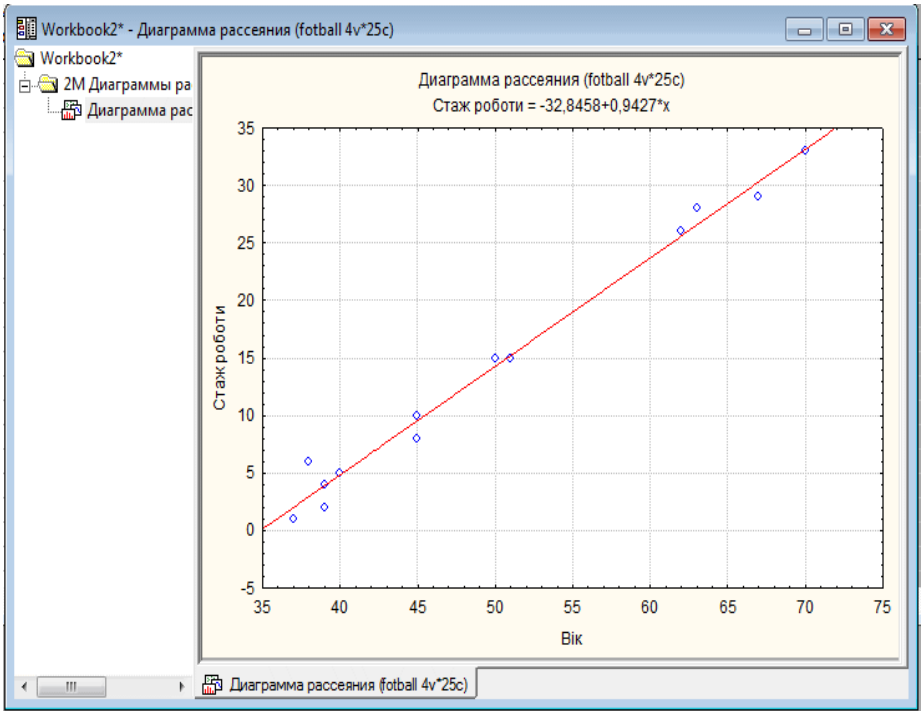
Після виконання будь-якої з цих дій з'явиться вікно "*2M Диаграммы рассеяния*" ("2D Scatterplots").

2. На вкладці *"Быстрый" ("Quick"*) вибрати змінні, за якими будується графік. Для цього слід натиснути кнопку **«Переменные» («Variables»)**, що спричиніть появу вікна "*Выберите переменные для диаграммы рассеяния*" ("Select Variables for Scatterplot"). У лівому списку вікна вибирається факторна (*незалежна*) змінна, а у правому – результативна, що залежить від змінної-фактора.

3. На цій же вкладці знаходиться група показників "Регрессия" (Regression bands), за допомогою яких задається рівень довірчого інтервалу для лінії регресії.

4. Закрити послідовно вікна натисканням **«ОК»**, після чого відкриється вікно робочої книги з побудованим графіком.

Червона лінія регресії графіку дає уявлення щодо лінійності зв'язку між змінними: якщо точки близько розташовані біля лінії, то можна говорити щодо наявності лінійної залежності між змінними.

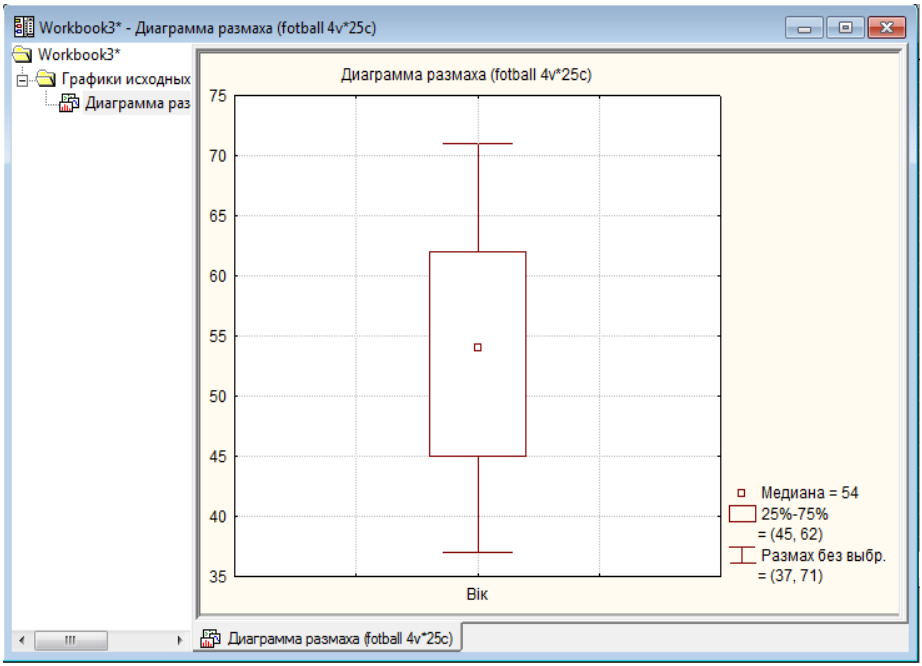


Діаграма розмаху

Таку діаграму ще називають *коробка з вусами, скринька з вусами (Box Plot, Box & Whisker Plot).*

Інформація графіка є дуже змістовною і корисною. У загальному випадку на ньому водночас графічно відображується кілька статистичних показників, що характеризують первинні

дані: мінімальне та максимальне значення, середня або медіана, перший квартиль (або 25 процентиль) та третій квартиль (75 процентиль). Отже такі діаграми не тільки відображують основні характеристики розподілу, але можуть бути використані також і для оцінки розмаху варіації та асиметрії. Основою діаграми є вертикальний або горизонтальний прямокутник, нижній бік якого (або лівий бік, якщо прямокутник розташовано горизонтально) є нижнім квартилем (Q_1), а верхній (правий) є верхнім квартилем (Q_3). Таким чином висота (або довжина) прямокутника дорівнює міжквартильному інтервалу (*IQR*). Невеличкий квадрат у площі прямокутника відображає середню арифметичну або медіану. Особливістю графіка є наявність "вусів" (*Whisker [wiske]*), якими є вертикальні або горизонтальні лінії, довжина яких відповідає вибраному значенню показника розкидання даних (це може бути максимум і мінімум, стандартне відхилення, дисперсія, квартилі) або точності оцінки генеральних параметрів (стандартна похибка, довірчий інтервал).



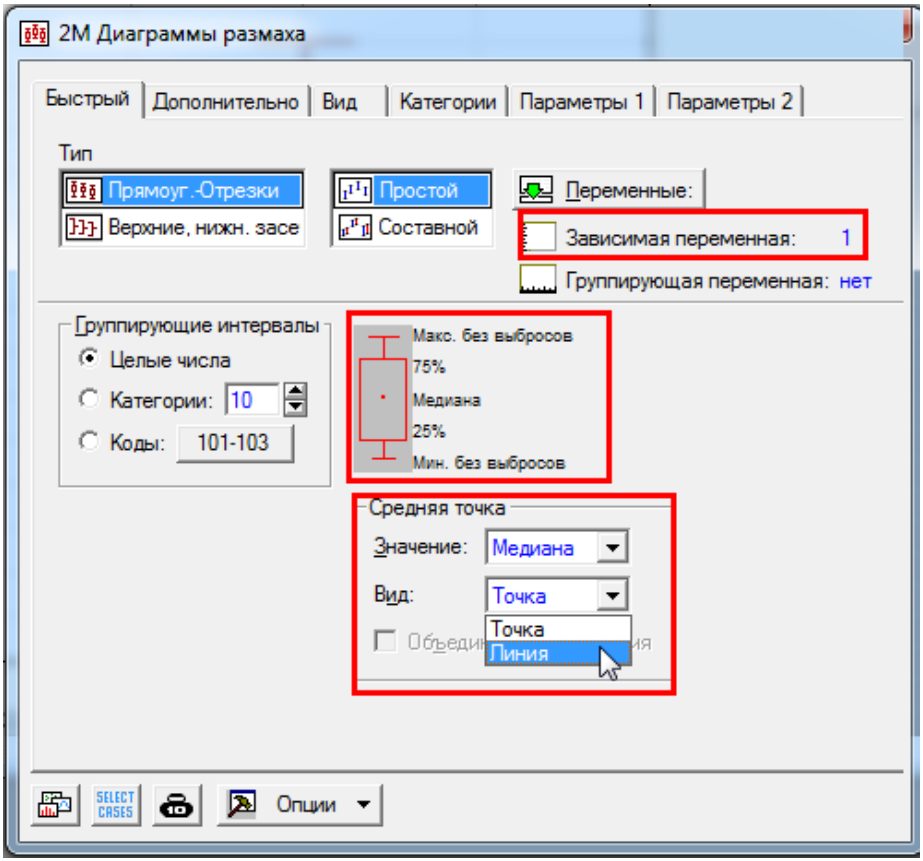
Діаграма використовується у багатьох статистичних методах, зокрема для візуалізації описових статистик, оцінки змін у часі або між різними групами, у кореляційному аналізу тощо. Відповідно її побудову можна здійснити у різних аналізах. Якщо ж будувати графік самостійно, то це здійснюється так:

1. Звернутися до модуля побудови діаграми розмаху, наприклад за командою **Графика ▸ 2M графики ▸ Диаграммы размаха…** (**Graphs ▸ 2D Graphs ▸ Box Plots…**). Після чого з'явиться вікно "*2M Диаграммы размаха*" ("2D Box Plots").

2. Визначити змінну, за якою будується графік, для чого на вкладці *"Быстрый" ("Quick"*) слід натиснути кнопку **«Переменные» («Variables»)** і у вікні "*Выберите переменные для диаграммы рассеяния*" ("Select Variables for Scatterplot") вибрати потрібну змінну, використовуючи для цього *залежну* змінну ("Зависимая переменная", "Depended variable"). Але досить часто визначаються дві змінні: залежна і змінна, *за якою здійснюється*

групування ("Группирующая переменная", Grouping variable). Такий підхід зокрема використовується, коли потрібно об’єднати однотипні значення змінної за групами. Наприклад, залежною змінною може бути заробітна плата співробітників, а змінною, за якою здійснюється групування – категорії співробітників: керівники, службовці, робітники.

3. На вкладці *"Быстрый" ("Quick"*) відображається невеличке зображення графіка із показником центру розподілу (середня арифметична (mean) або медіана (median) і тими значеннями показника розкидання даних, що будуть застосовані до "вусів" діаграми. У групі "Центральная точка" (Middle point) можна швидко змінити показники центру розподілу: зі списку "Значение" (Value) вибирається показник центру розподілу, а значення зі списку "Стиль" (Style) визначає, яким чином буде відображатися центр на графіку: у вигляді точки або лінії.



Вибір центру розподілу, а також інших елементі діаграми залежить від того, чи є він *нормальним*, а це в свою чергу впливає на перелік елементів діаграми.

Які саме значення слід вибирати демонструє наступна таблиця.

Таблиця 2

Значення графічних елементів діаграми розмаху

Графічний елемент	**Значення графічного елемента залежно від типу розподілу**	
	Нормальний	**Відмінний від нормального**
Вуси	Стандартна похибка / 95 % ДИ	Розмах / Non-Outlier Range
Коробка	Стандартне відхилення	Міжквартильний розмах
Центр	Середня арифметична	Медіана

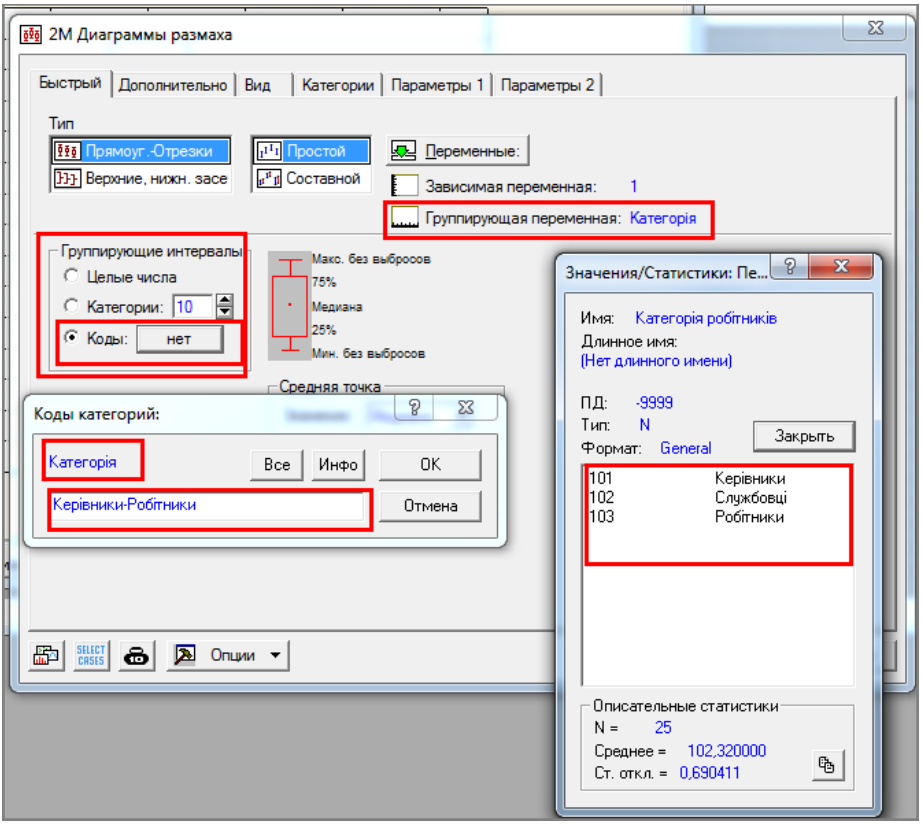
4. Змінити показник центру розподілу можна також на вкладці "*Дополнительно*" ("Advanced"). На цій вкладці також знаходяться настройки, за допомогою яких задаються інші елементи діаграми:

- Група "Размах" (Box) призначена для визначення основи прямокутника ("коробки") діаграми. За замовченням його нижній бік є 1-м квартилем (*Q1*), а верхній є 3-м квартилем (*Q3*), що визначається у списку "Значение" (Value) значенням "Процентили" (Percentiles) і числовим значенням "25" в полі "Коффициент" (Coefficient).
- Група "Усы" (Whisker) визначає статистичний показник, що буде задіяний на графіку у вигляді вусів. Він вибирається зі списку "Значение" (Value).
- Група "Выбросы" (Outliers) дозволяє відображати або ні так звані точки-викиди, тобто значення змінної, що суттєво відрізняються у більший або менший бік порівняно з іншими значеннями вибірки. Для відключення відображення викидів у списку "Выбросы" (Outliers) слід вибрати значення "Нет" (Off).

5. За наявності змінної, за якою здійснюється групування, можна обмежити кількість груп. Для цього на вкладці *"Быстрый" ("Quick")* у групі "Группирующие интервалы" (Grouping intervals) слід виконати такі дії:

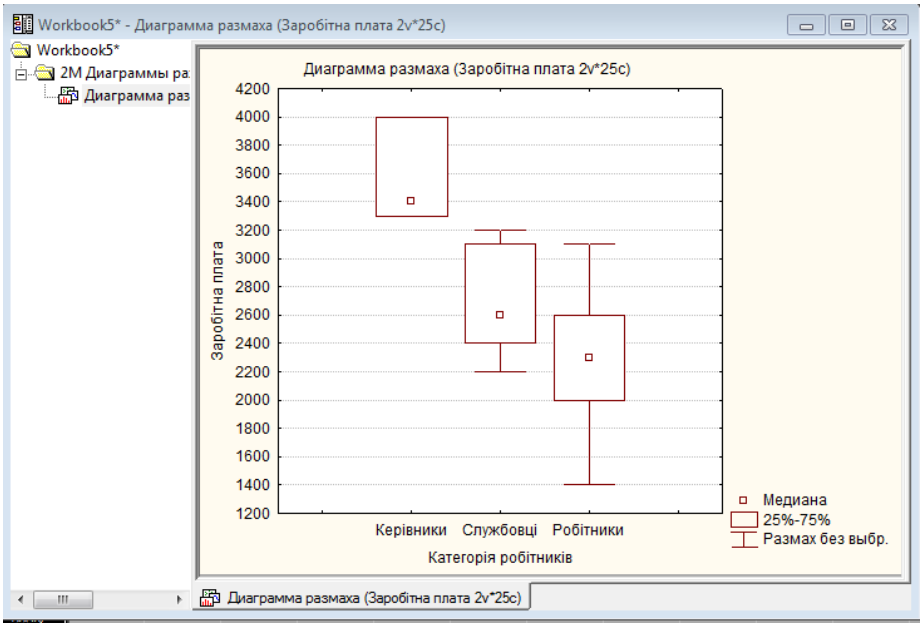
- встановити перемикач для групи "Группирующие интервалы" (Grouping intervals) в положення "Коды" (Codes);
- клацнути кнопку **«Нет»** (None), після чого з'явиться вікно "*Коды категорий*" (Category Codes);

- для вибору груп натисніть у ньому кнопку праворуч від перемикача "Коды", що спричиніть появу вікна з переліком груп;



- здійснити вибір потрібних груп, натискаючи на назві;
- послідовно закрити всі відкриті вікна натисканням кнопки **«ОК»**.

6. Натиснути кнопку **«ОК»**.



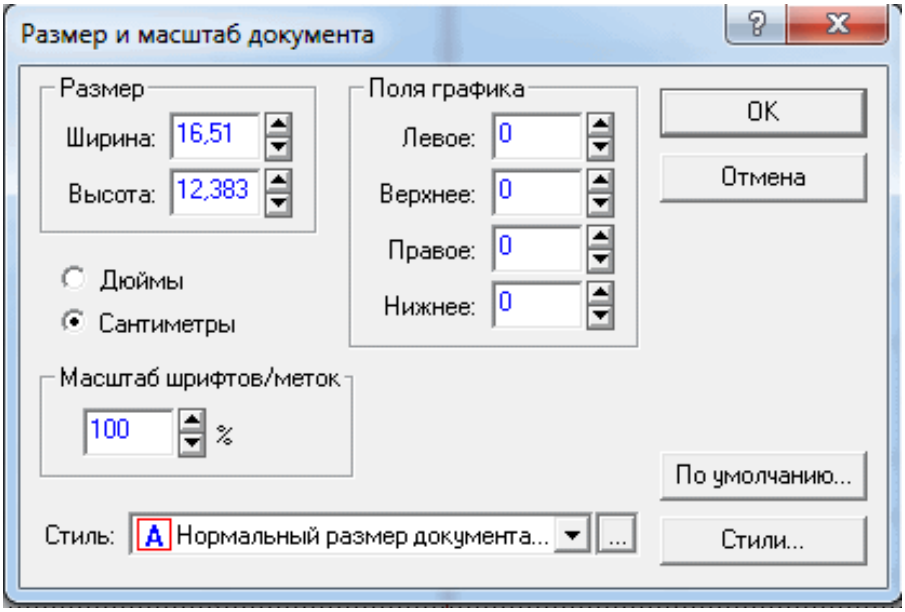
Інструменти роботи з графіками

Вікно з графіком має власну панель інструментів. Вона містить інструменти для рисування, настроювання параметрів і зовнішнього вигляду графіка, інструменти керування об'єктами у вікні, наприклад, зв'язування, вбудування графіків і графічних об'єктів.

Розглянемо найбільш вживані інструменти керування графіками.

Після створення графік відображається в режимі  "Растянуть график" (Display Graph fit to windows), коли розміри графіка підганяються під розміри вікна. Інший режим відображення "Исходный размер графика" (Display Graph of the actual size). За другим режимом розміри графіка можуть бути такими, що він не зможе повністю відобразитися у вікні.

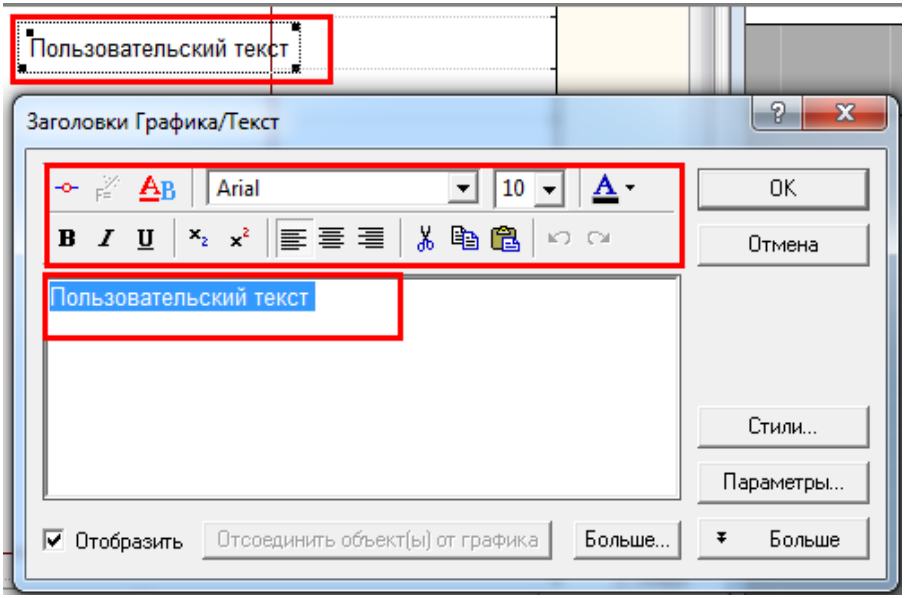
Визначити потрібні розміри вікна дозволяє інструмент настроювання розмірів  . Після вибору цього інструмента з'являється вікно "*Размер и масштаб документа*" ("Document Size and Scaling", Розміри і масштаб документа), в якому і задаються потрібні розміри. За замовчуванням розміри визначаються в дюймах, тому для визначення їх в метричній шкалі слід встановити перемикач одиниць виміру в положення "сантиметры" ("centimeters").



Користувач має можливість збільшити або зменшити будь-яку частину графіка, що здійснюється за допомогою інструментів збільшення / зменшення:  . Після вибору інструмента слід встановити курсор на потрібну область графіка і натиснути ліву кнопку миші.

Кілька інструментів дозволяють нанести на графік додаткову інформацію: текст, низку графічних примітивів – прямокутники, еліпси і навіть взагалі вставити інший об'єкт. Наприклад, дода-

вання до графіка тексту здійснюється за допомогою інструмента "Текст"  . Після вибору цього інструмента слід встановити курсор в поле графіка і натиснути кнопку миші, що призведе до появи на цьому місці текстового об'єкта "Пользовательский текст" ("Custom Text"), який надалі редагується звичайним чином із застосуванням стандартних інструментів форматування.



Редагування графіків

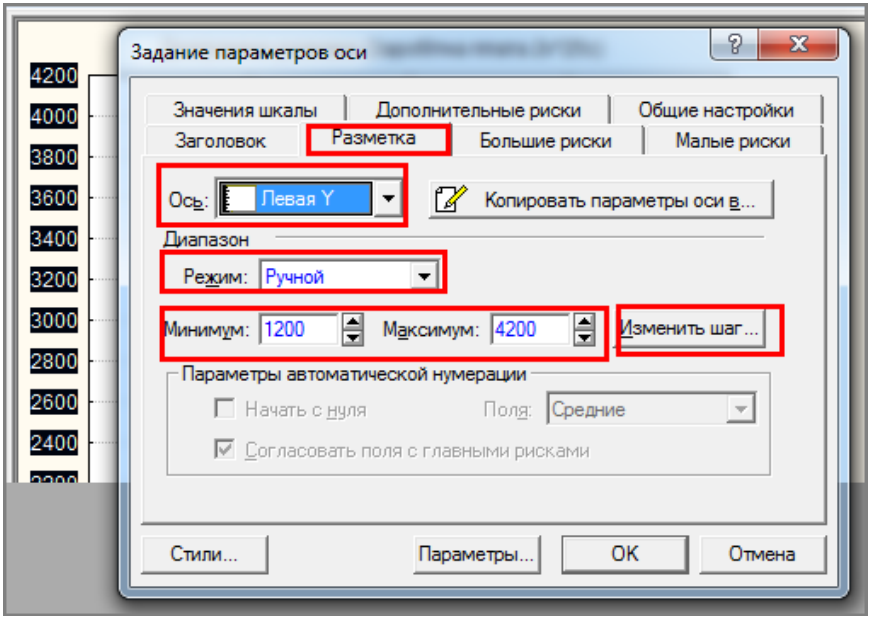
Побудованому графіку можна надати більш зручний вигляд шляхом його редагування його елементів. Редагуванню підлягають усі його елементи: загальна назва, заголовки осей і т. ін. Для вибору об'єкта застосовується інструмент "Покажчик"  . Він дозволяє виділити об'єкт простим натисканням на ньому. Подвійне натискання на об'єкті відкриває вікно його редагування. Перехід до редагування графіка можна здійснити і іншим шляхом. Для цього потрібно натиснути на об'єкті мишею і викликати контекст-

не меню. Пункт меню "Параметры" ("Properties", властивості) відкриває вікно з усіма властивостями об’єкта.

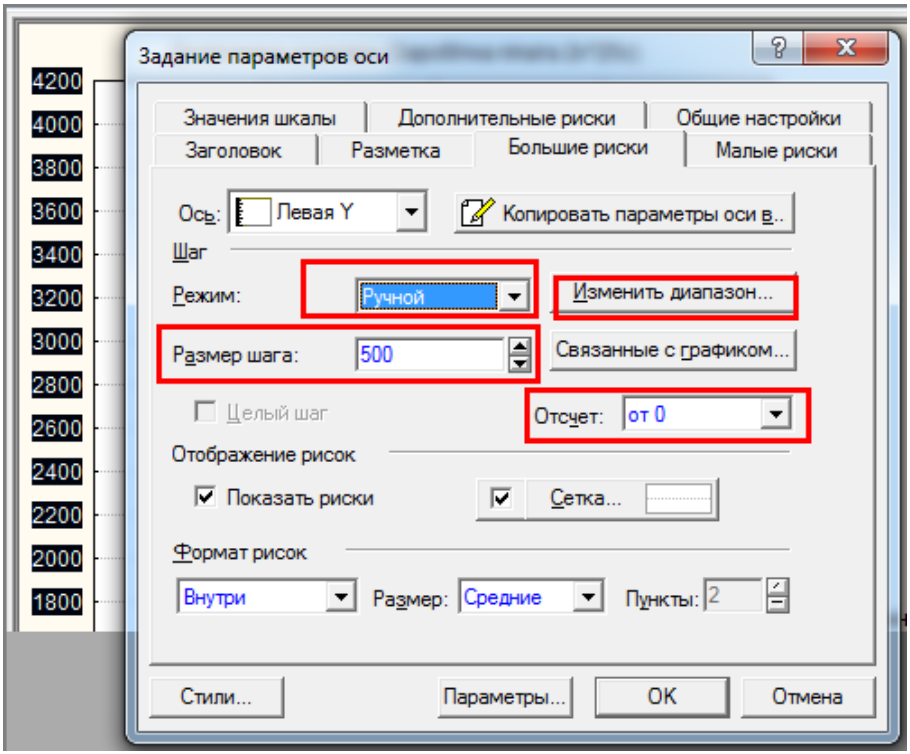
Щоб дістатися до властивостей усіх об’єктів графіка слід двічі натиснути на ньому. З’явиться вікно "*Задание параметров графика*" ("Graph Options"). Визначимо деякі дії редагування.

1. Змінити загальний заголовок графіка: вкладка "Заголовки Графика/Текст" ("Graph Titles/Text").
2. Змінити назви осей: вкладка "Ось: Заголовок" ("Axis: Title").
3. Змінити початкове значення шкали осі та вибрати новий числовий інтервал шкали (крок) вкладка "Ось: Разметка" ("Axis: Scaling") або викликати контекстне меню на осі і вибрати з нього пункт "Свойства оси" (Scaling).

- Для переходу в режим ручного редагування нумерації осі слід зі списку "Режим" ("Mode") вибрати елемент "Ручной" ("Manual", ручний), після чого стануть доступними поля "Минимум:" ("Minimum:") і "Максимум:" ("Maximum:"), в яких задаються початкове і кінцеве значення числової шкали.



- Для зміни величини кроку інтервалу слід натиснути кнопку **«Изменить шаг» («Edit Step»**, змінити крок). Автоматично відбудеться перехід на вкладку "Большие риски" ("Major Units"). Так саме зі списку "Режим" ("Mode") слід вибрати елемент "Ручной" ("Manual") і в полі "Размер шага" ("Step Size") задати нове значення кроку. Для повернення до режиму редагування числової шкали слід натиснути кнопку **«Изменить диапазон» («Edit Range»)**.



- За закінченням редагування натиснути кнопку **«ОК»**.

Редагування первинних даних графіків

Як правило, первинні дані не бувають "у чистому вигляді": серед них зустрічаються помилкові та аномальні значення, викиди, пропуски і т. ін., які спотворюють форму графіка. Проте програма має спеціальний засіб, за допомогою якого на графіку можна вилучати окремі дані. Для цього слід виконати команду **Формат ▸ Редактор данных графика (Format ▸ Graph Data Editor**, Редактор даних графіка). З'явиться вікно з даними всіх спостережень, у якому достатньо просто змінити значення даних або викликати на рядку з даними для певного спостереження контекстне меню і вибрати, наприклад "Вырезать" (Cut, Вирізати).

Ще простіше виконати дію вилучення даних на графіку можна візуально за допомогою інструмента "Пензель". Це здійснюється за таким алгоритмом:

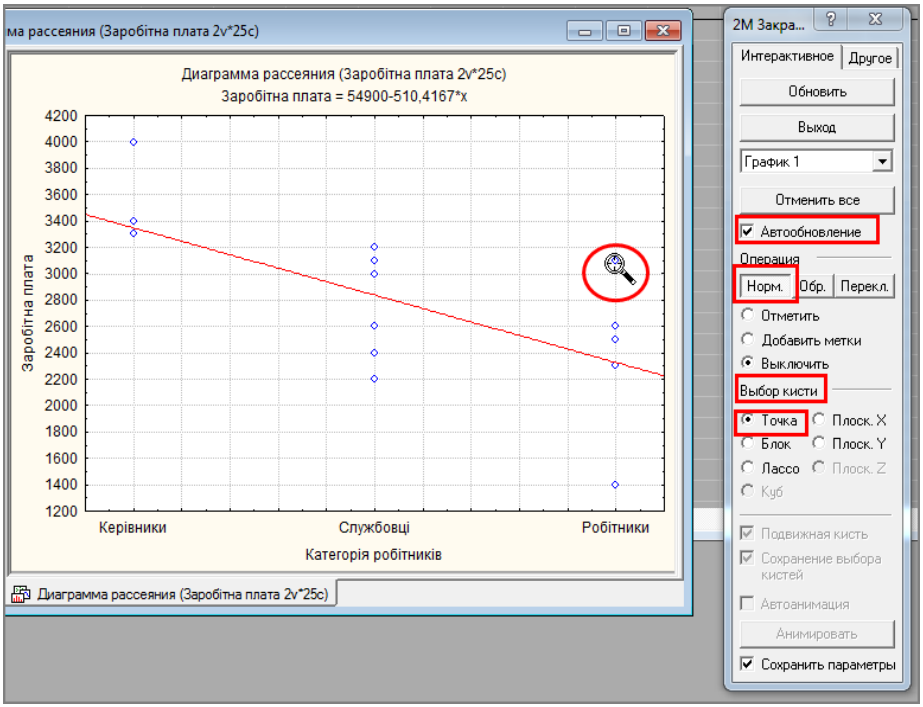
1. Виконати команду **Вид ▸ Закрашивание (View ▸ Brushing)** або натиснути кнопку  на панелі інструментів графіка. З'явиться панель "*2M Закрашивание*" ("Brushing 2D").

2. У групі полів "Операция" ("Action") повинна бути "утоплена" кнопка **«Норм.»** ("Normal"), а перемикач установлений в положення "Выключить" ("Turn OFF", вимкнути).

3. Для автоматичного відображення змін встановити позначку для поля-мітки "Автообновление" ("Auto Update", автоматичне оновлення).

4. У групі полів "Выбор кисти" ("Selection Brush", Вибір форми пензля) встановити перемикач в положення "Точка" ("Point"), якщо вилученню підлягають окремі точки, "Блок" ("Box") або "Лассо" ("Lasso"), якщо потрібно вилучити групу точок на прямокутній ділянці або ділянці довільної форми.

5. Навести курсор на точку (на графіку він набуває форму "прицілу") і натиснути ліву кнопку миші, що спричиніть вилучення точки, а графік змінить свій вигляд.



Для скасування змін слід натиснути кнопку **«Отменить все»** (**«Reset All»**, відмінити все).

Кореляційний аналіз

Якщо статистичні методи застосовують тільки для однієї змінної, то такі методи називають *одновимірні*. Проте одним з найважливіших завдань статистичного аналізу даних є виявлення й аналіз взаємозв'язків між змінними, для чого використовуються *багатовимірні* методи, зокрема, кореляційний і регресійний аналіз.

В економіці кореляційний аналіз використовується під час маркетингових обстежень, для аналізу діяльності господарської діяльності суб'єктів господарювання. В природничих дослідженнях метод з успіхом використовується під час обробки результатів

інтерференцій звукових та електромагнітних воли для визначення корисного сигналу, рівень якого набагато нижче рівня перешкод, що значно розширяє можливості зв'язку. Кореляційний аналіз займає відчутне місце під час вивчення медико-біологічних процесів, де він використовується для визначення зв'язку між зростом і масою людини, температурою тіла і частотою пульсу і т. ін.

Досить часто при проведенні досліджень (наприклад, маркетингових) саме вивчення кореляційних зв'язків є наступим етапом аналізу даних.

При проведенні аналізу даних особлива увага приділяється аналізу зв'язку між змінними з метою перевірки гіпотези щодо існування такого зв'язку.

Під час дослідження взаємозв'язків потрібно дати відповідь на три питання:

1. Чи існує залежність між змінними?
2. Яка інтенсивність цієї залежності?
3. Який напрям і характер цієї залежності?

Критерій кількісної оцінки залежності між змінними називають *коефіцієнтом кореляції*. Значення коефіцієнта кореляції змінюється в діапазоні від "-1" до "+1". Зрозуміло, що чим більше абсолютне значення коефіцієнта кореляції, тим більш щільним є зв'язок між змінними. При цьому якщо значення коефіцієнта кореляції є додатнім, то між ними існує пряме, однонаправлене співвідношення. За таким співвідношенням малі значення однієї змінної відповідають малим значенням іншої змінної, великі значення – великим. Якщо ж значення коефіцієнта є величиною від'ємною, то між ними існує зворотний, різноспрямований зв'язок. За такого зв'язку малим значенням однієї змінної відповідають великі значення іншої і навпаки.

Для орієнтовної інтерпретації значень коефіцієнта кореляції можна застосувати наступну таблицю:

Таблиця 3

Орієнтовна інтерпретація значень коефіцієнта кореляції

Значення коефіцієнта кореляції r	Інтерпретація
0 < r <= 0,2	Дуже слабка кореляція
0,2 < r <= 0,5	Слабка кореляція
0,5 < r <= 0,7	Середня кореляція
0,7 < r <= 0,9	Сильна кореляція
0,9 < r <= 1	Дуже сильна кореляція

Для змінних, що належать порядковій шкалі, застосовується *коефіцієнт Спірмена*, а для змінних, що належать до інтервальної шкали – *коефіцієнт кореляції Пірсона*. Слід мати на увазі, що кожну змінну, що належить до номінальної шкали і має дві категорії, можна розглядати як порядкову.

При цьому використання коефіцієнта кореляції Пірсона передбачає виконання двох обов'язкових умов:

1. розподіл значень обох змінних є нормальним;
2. зв'язок між змінними є лінійним.

Залежно від наявних даних для проведення кореляційного аналізу використовуються різні аналізи модуля "Основные статистики и таблицы" ("Bases Statistics/Tables").

Визначення парних взаємозв'язків відразу для кількох змінних вирішується шляхом побудови *матриці кореляції*. Для цього виконуються такі дії.

1. Завантажити модуль "Основные статистики и таблицы" ("Bases Statistics/Tables"), після чого з'явиться однойменне вікно.

2. У вікні вибору модуля вибрати пункт "Парные и частные корреляции" ("Correlation matrices").

3. Вибрати змінні, між якими слід визначати кореляцію. Це можна здійснити за допомогою двох інструментів.

- Інструмент "Квадратная матрица". При використанні цього інструмента коефіцієнти кореляції розраховуються попарно для всіх комбінацій відібраних змінних.

- Інструмент "Прямоугольная матрица". При використанні цього інструмента коефіцієнти кореляції розраховуються для всіх комбінацій змінних з першого та другого списків.

4. Натиснути кнопку **«Матрица парных корреляций» («Summary: Correlation matrices»)**, що спричиніть появу вікна з матрицею коефіцієнтів кореляції.

5. Водночас з розрахунком коефіцієнтів кореляції здійснюється оцінка їх статистичної значущості, тобто перевіряється нульова гіпотеза щодо наявності зв'язку між змінними.  Якщо значення коефіцієнта кореляції виділено червоним кольором, то це свідчить про те, що між змінними визнається існування зв'язку *для певного рівня значущості*.

6. Параметр "Уровень значимости для выделения" (Рівень значущості для виділення) знаходиться у вікні на вкладці "*Опции*" (Options) і за замовчуванням дорівнює "0,05". У статистиці при перевірці суттєвості зв'язків прийнято використовувати рівні значущості $\alpha = 0{,}05$ і $\alpha = 0{,}01$. Ці рівні означають, що за відсутності зв'язку між змінними лише в 5 чи 1 випадку із 100 теоретичне (критичне) значення коефіцієнта кореляції може перевищувати фактичне його значення.  Якщо фактичне значення коефіцієнта кореляції більше за критичне, то робиться висновок, що зв'язок між змінними є суттєвим. Зрозуміло, що зі збільшенням рівня значущості підвищується шанс визнання існування зв'язку між змінними.

Необхідно відмітити, що вибір певного рівня значущості, вище за який результати відкидаються як помилкові, є досить довільним. На практиці значення "0,05" є прийнятою межею статистичної значущості, проте слід пам'ятати, що цей рівень означає досить велику вірогідність помилки (5 %). Результати з рівнем значущості 0,01 зазвичай розглядаються як статистично значущі, а результати з рівнем 0,005 або 0,001 як високо значущі. Проте ця класифікація рівнів значущості довільна і ґрунтується лише на результатах практичного досвіду в певній галузі досліджень.

Алгоритм проведення кореляційного аналізу для залежних даних

Прикладом залежних даних можуть бути, наприклад, значення тієї самої змінної для різних періодів часу.

Проведення аналізу здійснюється так:

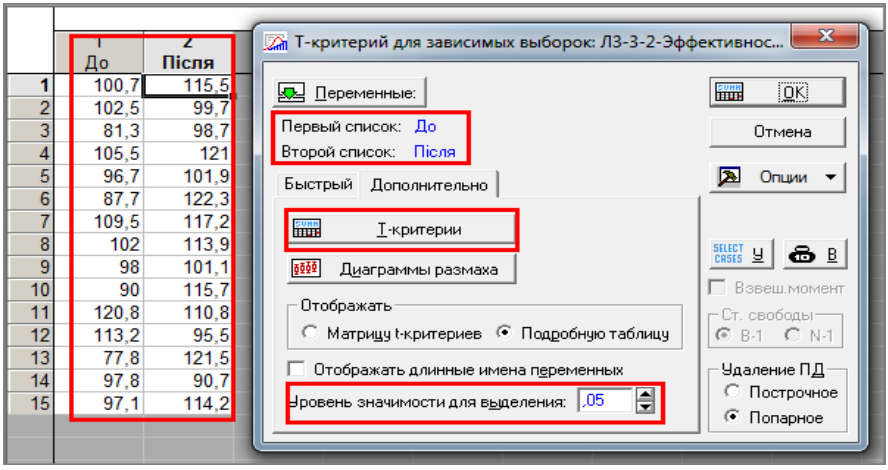
1. Завантажити модуль "Основные статистики и таблицы" ("Bases Statistics/Tables"), після чого з'явиться однойменне вікно.

2. У вікні вибору модуля вибрати аналіз "t-критерий для зависимых выборок" (t-критерій для залежних вибірок, t-test, dependent samples). З'явиться вікно "*T-критерий для зависимых выборок*" (T-test for Dependent Samples).

3. Натиснути кнопку **«Переменные» («Variables»)** і відібрати змінні для аналізу.

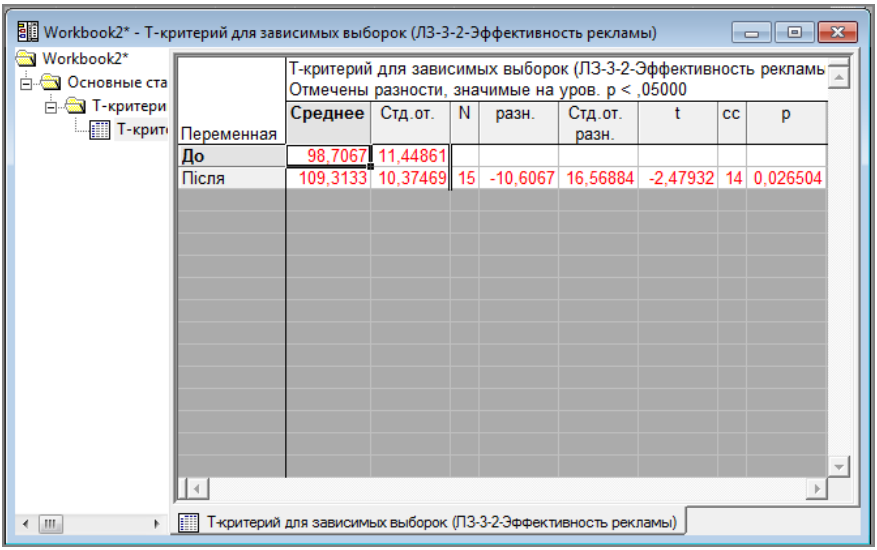
4. Закрити вікно відбору змінних, натиснувши **«ОК»**.

5. За необхідністю на вкладці "*Дополнительно*" ("Advanced") в полі "Уровень значимости для выделения" ("p-level for highlighting:", рівень значущості) можна встановити потрібний рівень значущості.



6. Натискання кнопки **«Т-критерии» («Summary: T-tests»)** ініціює розрахунок основних описових статистик і низки імовірнісних характеристик.

7. Натискання кнопки **«Диаграммы размаха» («Box & whisker plots»)** ініціює побудову діаграми розмаху.



Т-критерій Стьюдента призначено для оцінювання відмінностей середніх значень двох вибірок, розподіл яких відповідає нормальному закону. Одним з головних достоїнств критерію є широта його застосування. Він може використовуватися для співставлення середніх у залежних і незалежних вибірках різного обсягу. Разом із тим застосування t-критерію Стьюдента вимагає певних умов:

1. Вибірки, що порівнюються, повинні відповідати нормальному закону.

2. Вимірювання може бути проведено в шкалі інтервалів і відношень.

Алгоритм проведення кореляційного аналізу для незалежних даних

Незалежні дані – це дані, що формально не залежать одне від одного, наприклад, уподобання людини та його стать. Такі дані можуть бути представлені двома способами.

1. Як *незалежні вибірки*. За таким варіантом спостереження міститиме будь-які значення незалежної групувальної змінної. Наприклад, одне спостереження стосується одного респондента,

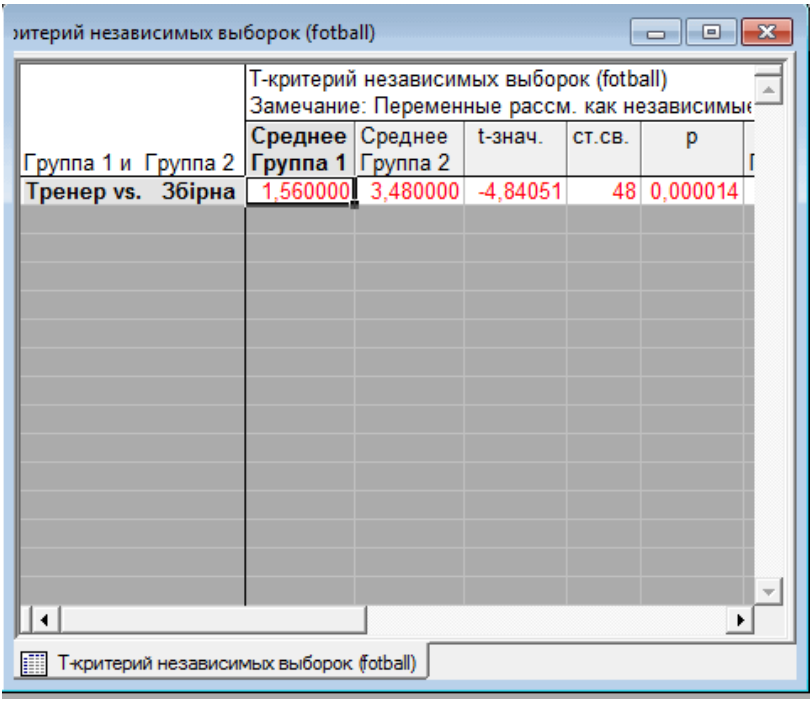
незалежно від його статі (чи це чоловік, чи – жінка). За цим варіантом групувальна змінна включається до аналізу для того, щоб визначити, до якої групи слід віднести певне спостереження.

2. Як *незалежні змінні*. За цим варіантом для кожної групи змінної створюється своє спостереження. Наприклад, одна для чоловічої статі, інша – для жіночої.

Проведення аналізу для незалежних вибірок здійснюється практично так саме, як і для залежних даних, з двома відмінностями.

1. У вікні "*Основные статистики и таблицы*" вибрати аналіз "t-критерий для независимых выборок" (t-критерій для незалежних вибірок). *З'явиться вікно "T-критерий для независимых выборок" ("T-test for Independent Samples").*

2. При виборі змінних для аналізу слід у першому списку вибрати залежну, а у другому – незалежну (групувальну) змінні, наприклад, уподобання і стать.



Регресійний аналіз

Досить часто вважають, що таке поняття "регресія" є важким для розуміння. Насправді, сутність явища регресії досить проста. Наприклад проводиться дослідження доходів телевізійних компаній. Серед чинників, що впливають на такі доходи, є, наприклад, надходження від реклами. А якщо телевізійна мережа є кабельною, то важливими чинниками будуть кількість абонентів та вартість підписки. Якщо розглядати дохід як наслідок (результат) дії цих чинників, то математично його можна подати у вигляді рівняння:

ДОХІД = b1 * Кількість_абонентів + b2 * Дохід_від_реклами + …

Отже математично регресію можна подати у вигляді рівняння, ліва частина якого містить результат (у даному випадку величину доходу телекомпанії), а права – кількість абонентів, помножене на коефіцієнт (абонентну плату), плюс доходи від реклами, помножені на коефіцієнт (доходів від реклами) і так далі:

ДОХІД = b1 * Кількість_абонентів + b2 * Дохід_від_реклами + …

Отже, маємо звичайну залежність результативної змінної від чинників. Після чого досить просто визначити числове значення доходу при зміні числового значення будь-якого чинника.

Типовим практичним завданням регресійного аналізу є виявлення залежностей між даними, наприклад, між ціною покупки акції і ціною її продажу, продуктивністю процесора та його вартістю, корисною площею житла та її вартістю, доходом та споживанням. В усіх цих прикладах перша змінна є фактором (чинником), а друга – залежною змінною (результатом). Першу позначають через x, а другу – через y. Як зазначалося раніше, дати кількісну характеристику залежності між змінними можна і у кореляційному аналізі. Але, на відміну від кореляційного аналізу, завдання регресійного аналізу полягає не тільки в тому, щоб з'ясувати залежність y від x , але досить часто ще й спрогнозувати значення y за певними значеннями x. Такі завдання розв'язуються шляхом побудови *регресійної моделі*.

Найпростішою регресійною моделлю є лінійна, проте і за її допомогою можна розв'язати багато практичних завдань.

Наведемо математичний опис лінійної регресійної моделі, в межах якої здійснюється дослідження залежності між змінними x та

y. Передбачається, що між змінними *x* та *y* існує залежність вигляду:

$$y(i) = b_0 + b_1 * x(i) + e(i), \quad (4)$$

де i – окремі спостереження вибірки (0<i<n);

b_0, b_1 – невідомі константи;

$e(i)$ – випадкові величини, що не спостерігалися, оскільки спостереження було проведено для *n* значень вибірки. Вони мають середню "0".

Іноді випадкові величини $e(i)$ ще називають похибками спостереження. Передбачається, що $e(i)$ не корелюють для різних вибірок.

Загальна постановка завдання полягає в тому, щоб на підставі спостережень пар *x* та *y* розв'язати такі основні питання:

1. дати найкращу оцінку параметрам моделі b_0, b_1;
2. побудувати довірчі інтервали для параметрів b_0, b_1;
3. перевірити гіпотезу про значущість регресії;
4. дати оцінку ступеня адекватності моделі.

Математичний розв'язок завдання

Розглянемо тільки першу частину завдання – оцінювання найкращим чином параметрів b_1, b_0.

Наше завдання за наявними даними спостереження побудувати пряму, що буде проходити максимально близько до усіх точок, що відповідають даним спостереження. У статистиці побудова такої прямої ґрунтується на однієї з властивостей середньої, згідно якої сума квадратів відхилень варіант від середньої арифметичної менша, ніж сума квадратів відхилень від будь-якої іншої величини:

$$\Sigma(x - \bar{x})^2 f = \min \quad (5)$$

Отже, передумовою побудови такої прямої є визначення середньої і окремих значень прямої за наведеною формулою. Відносно такої прямої кажуть, що вона побудована за методом найменших квадратів (МНК).

Наприклад, у результаті побудови прямої було отримане

рівняння y = 126,0427 + 188,9698*x*, яке називають *рівнянням регресії*. Оцінка вільного члена b_0 становить 126,0427, оцінка коефіцієнта b_1 (кута нахилу) – 188,9698. Ці оцінки є найкращими оцінками невідомих параметрів b_1, b_0, оскільки рівняння регресії максимально близько проходить до точок, побудованих за емпіричними даними. Такі оцінки називають оцінками, що побудовані *за методом найменших квадратів (МНК)* або просто *оцінками найменших квадратів*.

МНК можна застосовувати навіть без припущення щодо розподілу похибок. Проте, слід враховувати, що тільки за умови нормального розподілу оцінки параметрів моделі b_1, b_0 є оптимальними. Якщо розподіл відрізняється від нормального, то оптимальність може бути втрачена. Наприклад, це може бути якщо серед даних є такі, що суттєво відрізняються від більшості (так звані викиди), оскільки МНК чутливий до викидів.

В економіці регресійний аналіз широко використовується для прогнозування соціально-економічних показників. Так саме він є розповсюдженим у маркетингових дослідженнях, наприклад, вивчення попиту на товар залежно від ціни, рівня доходів населення, витрат на рекламу та інших чинників, вивчення залежності обсягу продукції від розміру інвестицій, технічного рівня устаткування, чисельності працівників тощо. У фінансовій економіці аналіз використовують для оцінки кредитоспроможності клієнтів. Регресійний аналіз займає відчутне місце в біомедицині, зокрема для розробки нормативних шкал і стандартів фізичного розвитку.

Алгоритм проведення регресійного аналізу

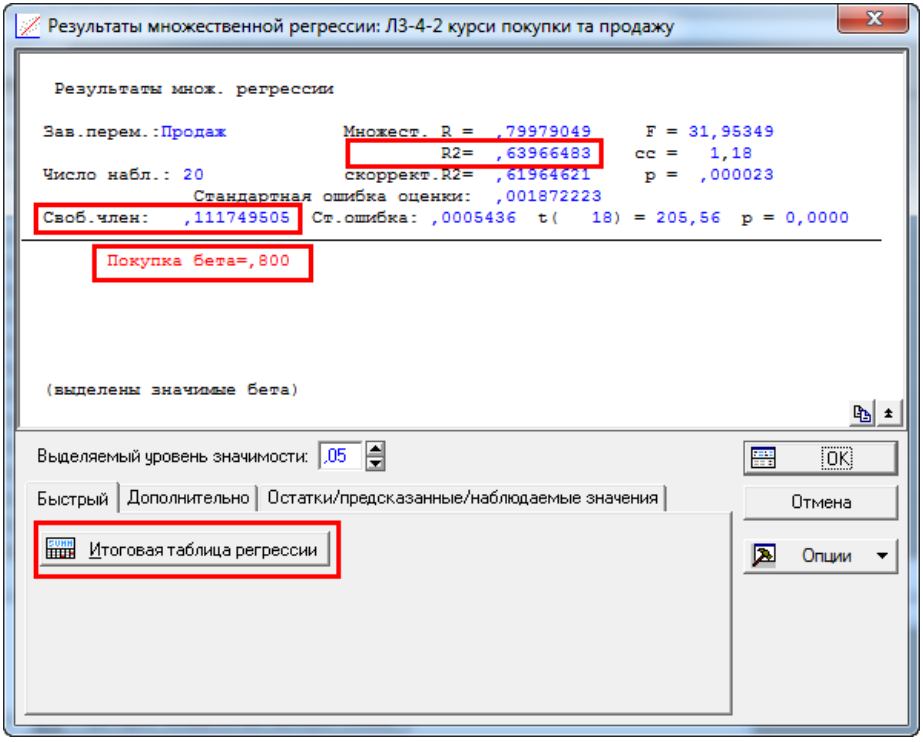
1. Виконати команду **Анализ ▸ Множественная регрессия (Statistics ▸ Multiply Regression)**. З’явиться стартове вікно модуля "Множественная регрессия" ("Multiple Linear Regression", Множинна лінійна регресія).

2. Вибрати змінні для аналізу. Для цього потрібно натиснути кнопку **«Переменные» («Variables»)**, після чого з’явиться вікно

для вибору. У правому списку вікна вибирається незалежна (факторна) змінна (независимая, Independent), у лівому списку – залежна (результативна) (зависимая, Dependent).

3. Натиснути **«ОК»** для повернення до стартового вікна модуля. В полях "Независимая" ("Independent") і "Зависимая" ("Dependent") відобразяться імена вибраних змінних.

4. Натиснути **«ОК»** у стартовому вікні модуля. З'явиться вікно "*Результаты множественной регрессии*" ("Multiple Regression Results:", Результати множинної регресії).



Верхня частина вікна містить такі показники:

1. Після тексту "Зависимая переменная" ("Dependent:") знаходиться ім'я залежної змінної ("Продаж");

2. Число наблюдений (Кількість спостережень, No. of

cases): 20.

3. Множест. (Multiple) R = (0,79979049). Коефіцієнт множинної кореляції.

4. R2. Квадрат коефіцієнта множинної кореляції, тобто ***коефіцієнт детермінації***.  Цей показник є найважливішим, оскільки він визначає частку розсіювання навколо середнього значення, що "пояснює" побудована регресія. Значення коефіцієнту детермінації знаходиться в межах від "0" до "1".  Чим більше його значення наближається до "1", тим більше регресія *пояснює розсіювання значень залежної змінної відносно вибіркової середньої*. Для нашого прикладу 0,64 – це досить високе значення, яке пояснює 64 % розсіяння значень залежної змінної (покупки) відносно середньої.

5. Скоррект. (adjusted) R2 = (0,61964621). Коефіцієнт детермінації, скоригований на число ступенів свободи. Він визначається за формулою

$$R_s^2 = (1 - R^2) \cdot \frac{n}{(n - p)} \qquad (6)$$

де n – кількість спостережень, p – кількість параметрів моделі, яка визначається як число незалежних змінних + 1, оскільки до моделі включено вільний член.

6. p = 0,000023. Рівень значущості.

7. Стандартная ошибка оценки (Standard error of estimate): (0,001872223). Стандартна похибка оцінювання, яка є мірою розсіювання значень спостережень відносно лінії регресії.

8. Свободный член (Intercept): (0,111749505). Значення вільного члена рівняння регресії, тобто значення коефіцієнта b_0 в рівнянні регресії.

9. Ст. ошибка (Std.Error): (0,0005436). Стандартна похибка оцінювання вільного члена.

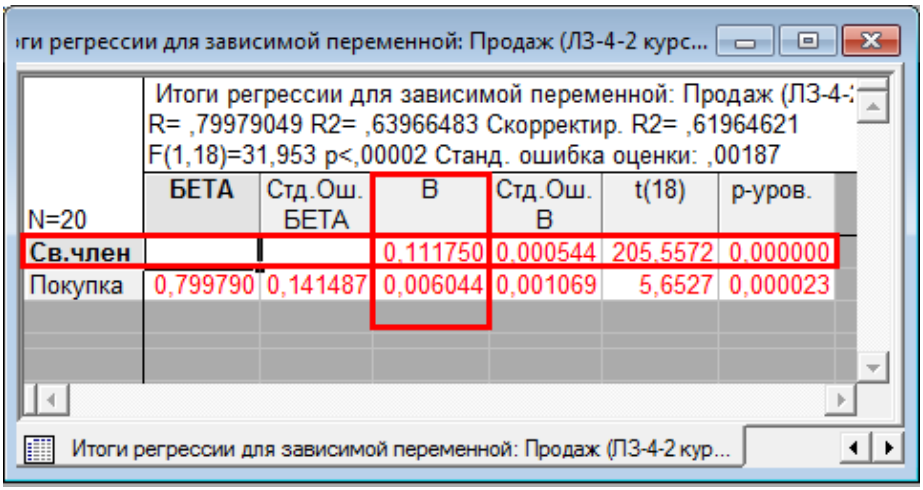
Середня частина інформаційного вікна містить стандартизований коефіцієнт регресії – *бета*, що був би отриманий у випадку

стандартизації змінних, тобто за таким перетворенням, коли середні змінних дорівнювали б "0", а стандартні відхилення – "1".

Розрахунок бета дозволяє оцінити, в який ступені значення залежної змінної описуються незалежними змінними. Цей показник є корисним особливо тоді, коли є кілька незалежних змінних з різними одиницями виміру. У цьому випадку бета відображає питомий внесок кожної незалежної змінної у варіацію залежної змінної. Якщо незалежна змінна тільки одна, то коефіцієнт бета співпадає зі значенням коефіцієнта множинної кореляції.

Якщо стандартизований коефіцієнт регресії виділений червоним кольором, то регресія є значущою.

Нижня частина вікна "*Результаты множественной регрессии*" містить кілька інструментів, за допомогою яких можна деталізувати результати. Так, після натискання кнопки «**Итоговая таблица регрессии**» (**«Summary: Regression results»**) відбувається розрахунок підсумкових результатів оцінювання регресійної моделі.



Стовпці таблиці результатів містять такі показники:

1. БЕТА: стандартизований коефіцієнт рівняння регресії.
2. Стд. Ош. БЕТА (Std.Err. of Beta, стандартна похибка Бета).
3.  Коефіцієнти рівняння регресії: клітинка першого рядка "Свободный член" (Intercept) містить вільний член рівняння регресії, інші – коефіцієнти b_i при незалежних змінних.
4. Стд. Ош. B (Std.Err. of B, Стандартні похибки) для коефіцієнтів рівняння регресії.
5. Значення t-критерія Стьюдента (t(кількість_ступенів_свободи)). Це значення використовується для перевірки нульової гіпотези про те, що коефіцієнти рівняння дорівнюють "0".
6. p-уровень (p-value, p-рівень). Ймовірність похибки для нульової гіпотези.

Оцінка адекватності моделі

Важливим елементом проведення регресійного аналізу є *оцінка адекватності моделі*: після того, як адекватність моделі доведена, її з високою ймовірністю можна використовувати на практиці для прогнозування.  Аналіз адекватності моделі базується на аналізі залишків. *Залишки* – це різниці між фактичними (емпіричними) значеннями спостереження і теоретичними, розрахованими за моделлю. Відповідно за фактичними даними будується ***емпірична*** крива розподілу, а за теоретичними – ***теоретична***. ***Теоретична*** крива розподілу відображає закономірність певного типу розподілу в чистому вигляді, тобто в тому випадку, коли на розподіл не впливають випадкові причини.

Аналіз оцінки адекватності моделі складається з двох етапів:

1. Залишки перевіряються на нормальність їх розподілу.
2. Дисперсія залишків повинна залишатися незмінною на всьому діапазоні значень змінних.

Алгоритм перевірки залишків на нормальність їх розподілу може бути таким:

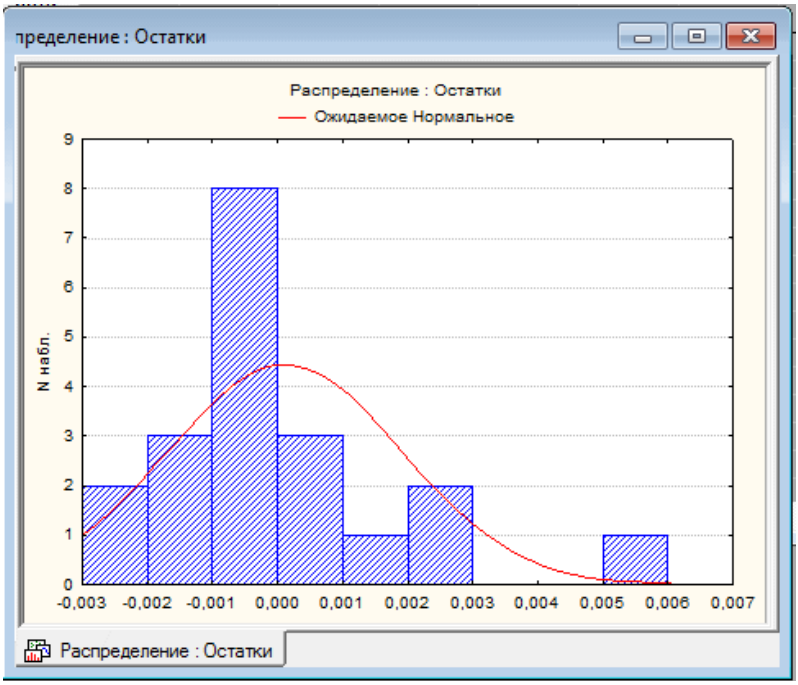
1. У вікні "*Результаты множественной регрессии*" (Результати множинної регресії) перейти на вкладку "*Остатки / предсказан-*

ные/наблюдаемые значения" ("Residuals/assumptions/ prediction", Залишки – Припущення – Прогнозування).

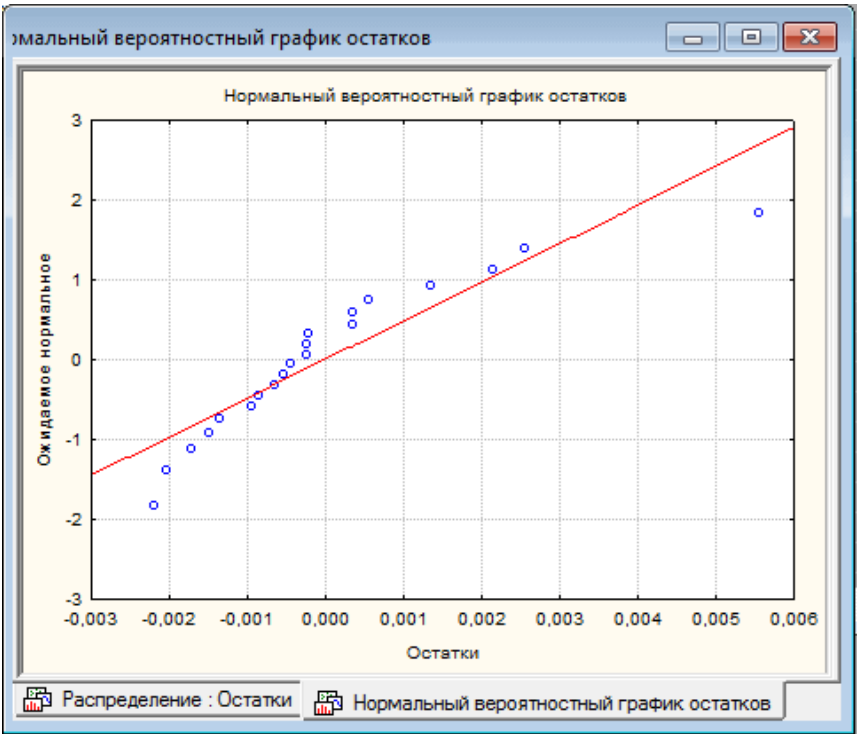
2. Натиснути кнопку **«Анализ остатков» («Perform residual analysis»**, провести аналіз залишків). З'явиться вікно "*Анализ остатков*" Residual Analysis).

3. Оскільки критерієм адекватності моделі можна вважати нормальність залишків, то гістограма розподілу залишків повинна бути наближена до графіка нормального розподілу. Для перевірки цього у вікні "*Анализ остатков*" слід перейти на вкладку "*Остатки*" ("Residuals", залишки). У групі полів "Тип остатков" ("Type of residual", Тип залишків) встановити перемикач у положення "Исходные" ("Raw residuals", по рядку) і натиснути кнопку **«Гистограмма остатков» («Histogram of residuals»,** Гістограма залишків).

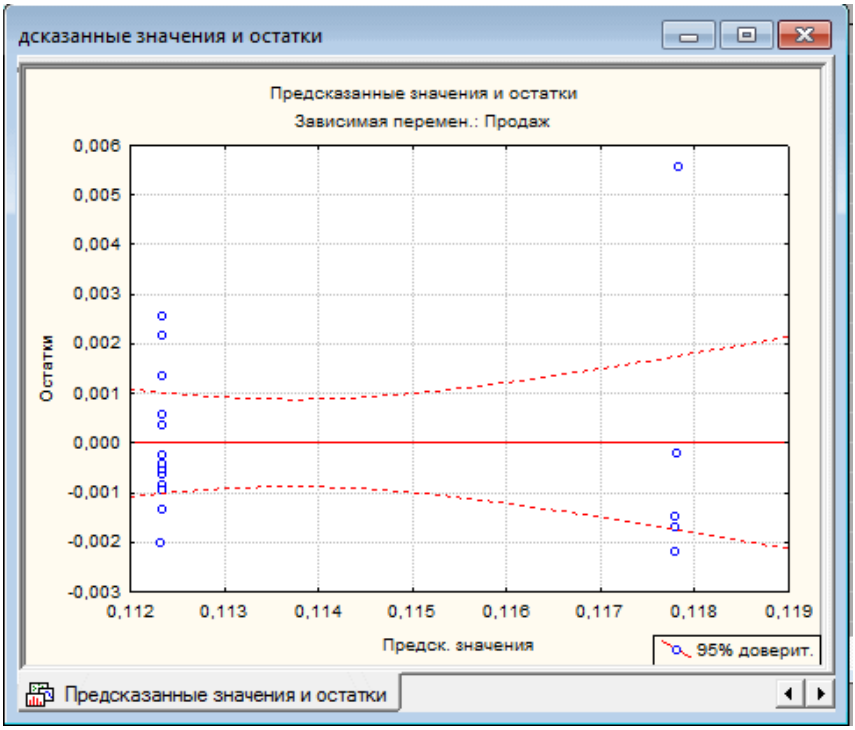
4. З'явиться гістограма розподілу залишків.  Якщо вона наближається до графіка нормального розподілу, то це й вказує на адекватність моделі.



За іншим варіантом для перевірки нормальності розподілу залишків у вікні "*Анализ остатков*" (Residual Analysis) на вкладці "*Быстрый*" ("Quick") слід натиснути кнопку **«Нормальный график остатков»** (**«Normal plot of residuals»**) і проаналізувати побудований графік. Якщо точки-залишки графіка достатньо близько розташовані біля (червоної) теоретичної прямої, то це дозволяє зробити висновок про нормальність розподілу залишків.



Другою умовою адекватності моделі є перевірка умови, що дисперсія залишків є незмінною на всьому діапазоні значень змінних. Для цього слід перейти на вкладку *"Диаграммы рассеяния"* ("Scatterplots") і натиснути кнопку **«Предсказанные и остатки»** (**«Predicted vs. residual»**, Теоретичні та залишки) для побудови графіку залежності значень залишків від теоретичних значень залежної змінної. Якщо умова виконується, то точки на графіку розташовуються хаотично, не відбиваючи жодної закономірності. Якщо ж точки розташовані близько до прямої або їх розташування має тенденцію, наприклад розкидання точок збільшується справа-наліво, то лінійний регресійний аналіз застосовувати неможна.



Для побудови графіку, що містить дані спостереження (фактичні дані) і теоретичну криву розподілу, побудовану за моделлю, слід перейти на вкладку "*Диаграммы рассеяния*" ("Scatterplots") і натиснути кнопку **«Предсказанные и наблюдаемые» («Predicted vs. observed»**.

Часто – якщо залишки не є нормальними – здійснюють перетворення залежних та незалежних змінних, наприклад, логарифмічне перетворення залежних змінних або обчислення з них квадратного кореня.

Як правило, первинні дані не бувають "у чистому вигляді": серед них зустрічаються помилкові та аномальні значення, викиди, пропуски і т. ін., які можуть суттєво вплинути на побудову моделі. Тому перед побудовою регресійної моделі слід вилучити такі дані дозволяє, що можна здійснити за допомогою редактора даних

графіка (див. розділ "Статистичні графіки").

Використання регресійної моделі для прогнозування

Для прогнозування значень залежної змінної слід виконати такі дії.

1. У вікні "*Результаты множественной регрессии*" перейти на вкладку "Остатки /предсказанные/ наблюдаемые значения" і натиснути кнопку **«Предсказать зависимую переменную» («Predict dependent variable»**, Прогноз залежної змінної), що спричиніть появу вікна "*Задайте значения независимых переменных*" ("Specify values for indep. vars", Введіть значення для незалежних змінних).

2. У полі з назвою незалежної змінної ввести прогнозне значення. Якщо незалежних змінних буде кілька, то буде відповідно і кілька таких полів.

3. Натиснути **«ОК»**. З'явиться вікно "*Предск(азанное) значения*" ("Predicting Values for", Прогнозоване значення), в якому рядок "Предсказ." (Predicted) містить прогнозне значення.

Предск.значения для (ЛЗ-4-2 курси покупки та продажу акцій на біржі)
перемен.: Продаж 1

Переменная	B-Веса	Значение	B-Веса * знач.
Покупка	0,006044	1,010000	0,006104
Св.член			0,111750
Предсказ.			0,117854
-95,0%ДП			0,116080
+95,0%ДП			0,119627

Дисперсійний аналіз

Дисперсійний аналіз є одним з методів математичної статистики, спрямованим на пошук залежностей експериментальних даних шляхом дослідження значущості різниць середніх значень. При цьому оброблюються кілька вибірок, що об'єднані в єдиній таблиці. На відміну від t-критерія, цей метод дозволяє порівню-

вати середні значення трьох і більше груп.

Метод розроблено біологом Р. Фішером в 1925 р. саме для аналізу результатів експериментальних досліджень у рослинництві. Іншими сферами застосування дисперсійного аналізу є експерименти у медицині, педагогіці, психології. Так у практичній діяльності лікарів під час проведення досліджень виникає необхідність встановлення впливу факторів на результати вивчення стану здоров'я населення, при оцінці професійної діяльності, ефективності нововведень. Його також достатньо широко використовують в економіці, наприклад під час вивчення оцінки впливу різнорідної сировини на якість продукції, впливу кількості добрив на урожайність сільськогосподарської продукції.

Сутністю метода є вивчення впливу одного або кількох незалежних чинників (факторів) на залежну (результативну) змінну. Залежні змінні подаються у вигляді шкал. Незалежні змінні є категоріальними, тобто відбивають групову причетність і можуть мати дві або більше градації (рівня). Прикладами незалежної змінної x_i може бути стать ($x2_i$ – жіноча, $x1_i$ – чоловіча), тип експериментальної групи (контрольна, експериментальна) тощо.

У літературі для дисперсійного аналізу залежно від кількості включених в нього факторів застосовують назву ANOVA (від англ. *ANalysis Of Variance*), якщо здійснюється однофакторний аналіз і MANOVA – для багатофакторного. Однофакторний дисперсійний аналіз використовується в тих випадках, коли є три або більше незалежні вибірки, отримані з той самої генеральної сукупності шляхом зміни якого-небудь незалежного чинника, для якої немає кількісних вимірів. Для цих вибірок припускають, що вони мають різні вибіркові середні і однакові вибіркові дисперсії. Необхідно дати відповідь на питання, чи здійснює цей чинник істотний вплив на розкидання вибіркових середніх або це розкидання є наслідком випадковостей, викликаних невеликими обсягами вибірок. Іншими словами, якщо вибірки належать тій самий генеральній сукупності, то розкидання даних між вибірками (між групами) має бути не більше, ніж розкидання даних усередині цих вибірок (усередині груп).

Нульова гіпотеза для дисперсійного аналізу формулюється як відсутність розбіжностей між груповими середніми результативної змінної:

$$H_0: \bar{x}_1 = \bar{x}_2 = \cdots \bar{x}_m \quad (7)$$

Перевірка цієї гіпотези ґрунтується на декомпозиції варіації результативної змінної за джерелами формування.

Загальна варіація результативної змінної ознаки розкладається на дві складові:

1. Міжгрупову варіацію, зумовлену дією фактора, покладеного в основу групування.

2. Внутрішньогрупову, випадкову варіацію.

Основну тотожність однофакторного дисперсійного аналізу можна подати як взаємозв'язок між сумами квадратів відхилень:

$$Q = Q_B + Q_W \quad (8)$$

де

$$Q = \sum_{1}^{m} \sum_{1}^{n_j} (x_{ij} - \bar{x})^2 \quad (9)$$

– сума квадратів відхилень окремих спостережень x_{ij} від загальної середньої $\bar{x}$;

$$Q_B = \sum_{1}^{m} n_j \left(\bar{x}_j - \bar{x}\right)^2 \quad (10)$$

– сума квадратів відхилень групових середніх $\bar{x}_j$ від загальної середньої $\bar{x}$;

$$Q_W = \sum_{1}^{m} \sum_{1}^{n_j} (x_{ij} - \bar{x}_j)^2 \quad (11)$$

– сума квадратів відхилень окремих спостережень x_{ij} всередині груп від групових середніх;

n_j – кількість спостережень у j-й групі; m – кількість груп (вибірок); $n = n_j m$ – загальна кількість спостережень.

На основі сум квадратів відхилень розраховуються три оцінки

дисперсій за джерелами варіацій:

загальна:

$$S^2 = \frac{Q}{n-1} \tag{12}$$

міжгрупова:

$$S_B^2 = \frac{Q_B}{m-1}; \tag{13}$$

внутрішньогрупова:

$$S_W^2 = \frac{Q_W}{n-m} \tag{14}$$

Знаменники оцінок дисперсії є ступенями свободи відповідних джерел варіації. Очевидно, вони співвідносяться так само, як суми квадратів відхилень:

$$(n-1) = (m-1) + (n-m) \tag{15}$$

Перевірка нульової гіпотези в однофакторному дисперсійному аналізі ґрунтується на співвідношенні міжгрупової і внутрішньогрупової варіації (в розрахунку на один ступінь свободи). F-тест показує, в скільки разів оцінка міжгрупової варіації перевищує внутрішньогрупову:

$$\mathrm{F} = \frac{\mathrm{S}_\mathrm{B}^2}{\mathrm{S}_\mathrm{W}^2} \tag{16}$$

Критичні значення F-критерію для рівня істотності α визначаються співвідношенням числа ступенів свободи чисельника $(m-1)$ і знаменника $(n-m)$. Процедура тестування стандартна: якщо $F > F_{1-\alpha}(\mathrm{m}-1;\mathrm{n}-\mathrm{m})$, то нульову гіпотезу відхиляють, а якщо $F < F_{1-\alpha}(\mathrm{m}-1;\mathrm{n}-\mathrm{m})$, то підстав для відхилення нульової гіпотези немає.

 Недоліком однофакторного дисперсійного аналізу є неможливість виявлення вибірок, що відрізняються від інших. З цією метою необхідно використовувати метод Шефе або проводити парні порівняння вибірок.

Алгоритм проведення дисперсійного аналізу

1. Виконати команду **Анализ ▸ Дисперсионный анализ (Statistics ▸ ANOVA)**. З'явиться вікно "*Общий ДА*" ("General ANOVA/MANOVA", Загальний дисперсійний аналіз).

2. Вибрати зі списку "Вид анализа" ("Type of analysis") пункт "Однофакторный ДА" ("One-way ANOVA").

3. Натиснути **«ОК»**. З'явиться вікно "*Однофакторный ДА*".

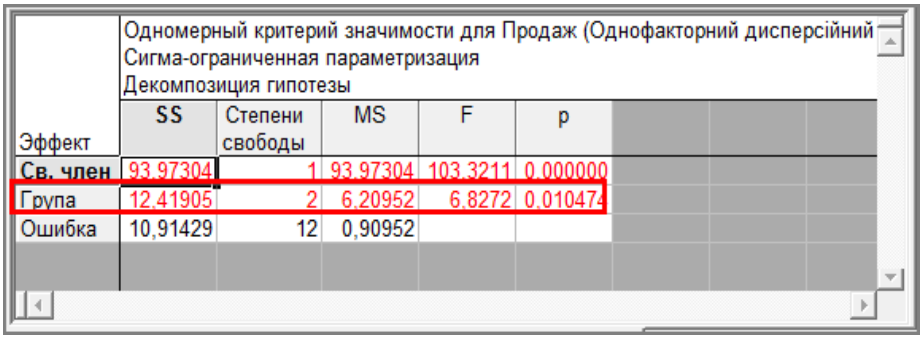
4. Натиснути кнопку **«Переменные» («Variables»)** і вибрати залежну (dependent) і незалежну (predictor) змінні у вікні "*Выберите зависимые переменные и категориальный предиктор*" ("Select dependent variables and a categorical predictor (factor)"). *Предиктор* – це незалежна змінна. Якщо її значення поєднуються у групи (категорії), то вона називається *категоріальною*, а якщо ні – то *неперервною*.

5. Натиснути **«ОК»**. З'явиться вікно "*Результаты анализа*" ("ANOVA Results").

6. Натиснути кнопку **«Все эффекты» («All effects»)**. З'явиться вікно "*Одномерный критерий значимости…*" для вибраної залежної змінної.

7. У вікні з результатами дисперсійного аналізу:

- рядок з назвою незалежної змінної відображає міжгрупову варіацію;
- рядок "Ошибка" ("Error") відображає внутрішньогрупову варіацію.
- Якщо значення у рядку з назвою незалежної змінної виділені червоним кольором, то це свідчить про те, що розбіжності між груповими середніми є істотними, тобто нульова гіпотеза про відсутність розбіжностей між груповими середніми результативної змінної відхиляється. Рядок з назвою незалежної змінної у колонці "F" містить значення F-критерію Фішера. У рядках з назвою незалежної змінної і "Ошибка" ("Error") відображаються значення ступенів свободи (Degr. of freedom), що надалі можна використати для розрахунку критерію Фішера.



Одномерный критерий значимости для Продаж (Однофакторний дисперсійний
Сигма-ограниченная параметризация
Декомпозиция гипотезы

Эффект	SS	Степени свободы	MS	F	p
Св. член	93,97304	1	93,97304	103,3211	0,000000
Група	12,41905	2	6,20952	6,8272	0,010474
Ошибка	10,91429	12	0,90952		

Додатковий аналіз визначення впливу незалежної змінної на залежну здійснюється за допомогою F-критерію Фішера:

1. За командою **Анализ ▸ Вероятностный калькулятор ▸ Распределения…** завантажити ймовірнісний калькулятор.

2. У списку "Распределение" (Розподіл) вибрати "F (Фишера)".

3. У полях "ст. св. 1" і "ст. св. 2" ввести значення ступенів свободи 1 і 2, обчисливши їх відповідно за формулами $(m-1)$ і $(n-m)$, де n – кількість спостережень, m – кількість вибірок.

4. У полі "p" задати ймовірність. Оскільки зазвичай для суспільно-економічних явищ приймають α=0,05 або α=0,01, то ввести у цьому полі значення "0,95" або "0,99".

5. Натиснути кнопку **«Вычислить»** (**«Compute»**, Обчислити).

6. Порівняти обчислене значення F-критерію зі значенням, що було розраховане під час проведення дисперсійного аналізу.

Дискримінантний аналіз

Дискримінантний аналіз – це один з методів багатовимірного статистичного аналізу. Його мета полягає в тому, щоб підставі вимірювання значень ознак об'єкту класифікувати його, тобто віднести до однієї з кількох груп (класів) деяким *оптимальним способом*. Під оптимальним способом розуміють або мінімум математичного очікування втрат або мінімальну ймовірність хибної класифікації.

Аналіз називають *багатовимірним*, оскільки вимірюються як мінімум дві ознаки.

Типові області застосування дискримінантного аналізу: медицина, економіка, геологія, контроль якості. Наприклад, в медицині об'єктом дослідження є пацієнт. За результатами його обстеження лікар приймає певні рішення, наприклад необхідність хірургічної операції. В економіці важливо віднести клієнта до певної групи при видаванні кредиту.

Математична постановка задачі

Є *n* об'єктів, які характеризуються *m* ознаками. Завдання полягає в тому, щоб за результатами спостережень віднести об'єкт до однієї з кількох груп (класів) $G_1,...,G_k, k \geq 2$. Іншими словами, слід сформулювати *вирішальне правило*, яке дозволить за результатами спостережень віднести новий об'єкт до конкретної групи. Кількість груп і до якої групи належіть той чи інший об'єкт заздалегідь відома.

Нехай X – простір значень вектора вимірювань. Вирішальне правило називається нерандомізованим, якщо простір X розподілений на *k* областей, які не перетинаються; при потраплянні вимірювання параметрів до *k-ї* області об'єкт відноситься до *k-ї* групи.

Вирішальне правило називається *рандомізованим*, якщо для кожного вектора спостережень *x* задана імовірність pi(x), з якою об'єкт належіть *i*-ї групі, *pi(x)>=0, p1(x) + ... + pk(x) = 1, i = 1,..., k.*

Зрозуміло, що при використанні вирішального правила виникають втрати *r(i, j)* під час неправильної класифікації об'єкта, тому вводять середні втрати, до яких призводить застосування правила, і прагнуть знайти правило, що мінімізує ці середні втрати.

Якщо значення втрат важко визначити чисельно, то при побудові оптимального правила використовують критерій мінімальної ймовірності хибної класифікації.

Під час дискримінантного аналізу можна задати апріорні ймовірності належності об'єкта до певного класу. На практиці ці ймовірності оцінюються за експериментальними даними.

 Для двох груп об’єктів дискримінантний аналіз тотожний множинній регресії (залежною змінною є номер групи).

В модулі дискримінантного аналізу реалізовано два методи: *стандартний* і *крок за кроком* (шляхом включення або виключення змінних). Вони аналогічні методам множинної регресії. Якщо груп дві, то за методом найменших квадратів будується лінія регресії (залежна змінна – групувальна, інші змінні – незалежні). Якщо груп більше, то і побудову умовно можна описати як процес дискримінації для 1-ї та 2-ї груп, потім – другої і третьої і т. д.

За методом "крок за кроком" модель будується поетапно. Для *метода включення* система на кожному кроці оцінює внесок у функцію дискримінації *ще не включених* до моделі змінних. Змінна з найбільшим внеском включається до моделі. Після цього система переходить до наступного кроку. Якщо застосовують *метод виключення*, то до моделі спочатку включаються усі змінні, а потім крок за кроком здійснюється їх послідовне виключення.

Алгоритм проведення дискримінантного аналізу

1. Звернутися до модуля дискримінантного аналізу за командою **Анализ ▸ Многомерный разведочный анализ ▸ Дискриминантный анализ**. Відкриється вікно "*Дискриминантный анализ*".

2. Визначити незалежні і групувальну (результатну) змінні.

3. Встановлення прапорця для поля-мітки "Дополнительные параметры" дозволяє деталізувати проведення аналізу.

4. Натиснути **«ОК»**. Якщо встановлений прапорець для дії "Дополнительные параметры", то з’явиться вікно "*Определение модели*" (Визначення моделі). Це вікно містить низку параметрів, які дозволяють деталізувати аналіз.

- На вкладці "*Быстрый"* можна визначити метод дискримінантного аналізу.
- На вкладці "*Дополнительно"* визначається таке.

4.1. Параметр "*Толерантность*" задає нижню межу толерантності: змінні, для яких значення толерантності менше цього значення, до моделі не включаються. Толе-

рантність розраховується за формулою $1 - R^2$ (*квадрат множинної кореляції*) змінної з незалежними змінними в моделі. Для методів "крок за кроком" моделі аналізуються на кожному кроці і кореляція обчислюється за включеними до моделі змінними.  Змінні з малим значенням толерантності можуть привести до помилок при обчисленні оберненої матриці. Очевидно, що якщо значення толерантності мале, то змінна має малу інформативність і включення її до моделі є недоцільним.

4.2. Параметри для методів "крок за кроком". "F-включить / исключить" – задають значення F-критерия для включення (або виключення) змінної до моделі.

- *Описательные*. Описові статистики.

5. Натиснути **«ОК»**. З'явиться вікно "*Результаты анализа дискриминантных функций*".

Аналіз здійснюється у вікні "*Результаты анализа дискриминантных функций*".

Верхня частина вікна містить, зокрема:

1. Кількість змінних в моделі.
2. Значення статистики F-критерію і рівень значущості (p).
3. Значення лямбди Уїлкса. Це значення містяться в інтервалі від "0" до "1".  Чим ближче воно наближається до "0", тим кращою вважається дискримінація.

Натискання кнопки **«Переменные в модели»** ініціює появу підсумкової таблиці аналізу даних.

Всі інструменти для аналізу і класифікації спостережень за групами знаходиться на вкладці "Классификация".

Правила класифікації ("Функции классификации")

За допомогою цих функцій можна визначити класифікаційні значення (мітки) для нових спостережень.  Ці значення можна трактувати як значення коефіцієнтів при відповідних змінних у рівнянні регресії, а значення константи – як його вільний член.

Отже, кожну групу можна описати своїм рівнянням. Для нового спостереження слід підставити значення його змінних до кожного рівняння і порівняти одержані класифікаційні значення.  Нове спостереження відносять до тієї групи (класу), для якої розраховане класифікаційне значення є мінімальним.

Правило "Квадраты расстояний Махаланобіса"

За допомогою цього методу будується таблиця, що містить квадрати відстані Махаланобіса від точок (спостережень) до центрів груп.  Нове спостереження відноситься до тієї групи (класу), для якої відстань Махаланобіса є мінімальною.

Символом "*" позначаються спостереження, що були неправильно класифіковані при використанні даного правила.

Правило "Апостериорные вероятности"

У вікні "*Априорные вероятности*" до аналізу для кожного спостереження задається ймовірність, з якою воно належіть до певного класу. Після виконання аналізу ці ймовірності можна заново обчислити й одержати апостеріорні ймовірності класифікації.

Натискання кнопки «**Апостериорные вероятности»** ініціює побудову таблиці з апостеріорними ймовірностями належності спостереження до певної групи.

Інтерпретація таблиці дуже проста. Перший стовпчик містить групу, інші – ймовірності віднесення спостереження до певної групи.  Спостереження відноситься до групи (класу) з максимальною апостеріорною ймовірністю.

Символом "*" позначаються спостереження, що були неправильно класифіковані при використанні даного правила.

Алгоритм визначення класу для нового спостереження

1. Додати в таблицю даних нове спостереження.
2. Визначити апостеріорну ймовірність належності спостереження до певної групи.
3. Визначити відстані Махаланобіса.

4. Визначити причетність до певної групи за одержаними результатами за наведеними вище правилами.

Кластерний аналіз

Кластерний аналіз об'єднує різні процедури, за допомогою яких дані поділяються на групи схожих об'єктів – *кластери*. ***Кластер*** – це об'єднання кількох однорідних елементів, яке може розглядатися як самостійна одиниця, що має певні властивості. В якості кластера можна розглядати об'єднання кількох пов'язаних родинними стосунками людей, тобто сім'ю. Кластером може бути клас, навчальна група і т. ін. Уперше термін "Кластерний аналіз" увів Трайон (Tryon) у 1939 р. Завданням кластерного аналізу є розподіл даних на групи (підмножини), що називаються кластерами, таким чином, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.

Найчастіше кластерний аналіз використовують в економіці, соціології, маркетингових дослідженнях, біології, медицині, археології.

Методи кластеризації поділяють на дві групи:

1. *Агломеративні* (агломерат – нагромадження). У таких методах (їх ще називають такими, що об'єднують) здійснюється послідовне об'єднання схожих об'єктів в кластери. Процес об'єднання може бути відображений на графіку у вигляді дендрограми (дерева об'єднання). Це досить зручне подання дозволяє наочно подати кластеризацію агломеративними алгоритмами.

2. *Ітеративні* (division – ділення, розподіл).

Даними для аналізу можуть бути безпосередньо об'єкти і їх параметри: об'єктами є спостереження, а їх параметрами – змінні. За іншим варіантом дані для аналізу подаються у вигляді матриці відстаней між об'єктами, у який на перетинанні рядка із номером *i* і стовпчика із номером *j* міститься відстань між *i*–м і *j*–м об'єктом.

 Якщо відстані не задані відразу, то реалізація агломеративних алгоритмів починається з обчислення відстаней між об'єктами.

Перехід від об'єктів до відстаней є важливим моментом. Відстань між об'єктами є однією із *мір подібності*. Суто інтуїтивно

зрозуміло, що чім менша відстань між об'єктами, тим більше вони схожі. Але під час розв'язування такої проблеми виникає питання вимірювання цієї відстані. У цьому разі зазвичай використовують *евклідову метрику*. Наприклад, якщо об'єкт описується двома параметрами, то його можна подати на площині точкою, а відстань між об'єктами обчислена за теоремою Піфагора. Тобто потрібно просто піднести до квадрату відстань за кожною координатою, просумувати їх і з одержаного результату добути корінь квадратний. За іншим варіантом відстань за кожною координатою не підносяться до квадрату, а беруться їх абсолютні значення. Такі відстані мають назву *манхетенскої відстані* або "*відстані міських кварталів*".

Але використання розглянутих вище метрик не вирішує усіх проблем. Наприклад, у соціологічних дослідженнях широко використовуються ознаки порядкової або номінальної шкали найменувань, що встановлює відношення подібності елементів, де відповіді мають вигляд на зразок "Краще – гірше". Зрозуміло, що такі дані практично не можливо подати точками на площині. У таких випадках використовується поняття *міри подібностей* об'єктів (відстань є однією з таких мір). У соціології часто ця міра подібності визначається безпосередньо. Наприклад, констатація співпадання відповідей призводить до міри подібності. Важливими мірами подібності, які часто використовують в соціальних науках, є статистичні коефіцієнти кореляції, наприклад, коефіцієнт Пірсона. Для бінарних даних просто обчислюють кількість параметрів, які співпадають у об'єктів, ділять одержане число на кількість параметрів і одержують міру подібності. Розраховані таким чином коефіцієнти називаються коефіцієнтами асоціативності.

STATISTICA має кілька мір подібності об'єктів:

- евклідова метрика;
- квадрат евклідової метрики;
- манхетенскої відстані або "відстані міських кварталів";
- метрика Чебишева;
- метрика Мінковського;
- коефіцієнт кореляції Пірсона;

- коефіцієнт со-зустрічаємості.

Вибір міри подібності об'єктів є процесом суб'єктивним і залежить від дослідника.

STATISTICA дозволяє здійснити кластеризацію використовуючи один з трьох методів:

1. Ієрархічна класифікація.
2. Двовходовє об'єднання.
3. Кластеризація методом К-середніх.

Перші два методи є *агломеративними*, третій – *ітераційним*.

 Ієрархічну класифікацію прийнято застосовувати якщо кількість кластерів заздалегідь невідомо, а метод К-середніх – якщо кількість кластерів дослідник визначає сам.

 Зазвичай застосуванню методів кластеризації передує процедура *стандартизації* даних. Ця процедура передбачає перетворення кожного даного у стовпці за формулою:

$$x' = \frac{x - \bar{x}}{\sigma} \tag{17}$$

де

$$\bar{x} = \frac{\Sigma x}{n} \text{ – середня величина,} \tag{18}$$

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \tag{19}$$

– середнє квадратичне відхилення.

Одержані в результаті цієї процедури дані мають нульову середню й одиничну дисперсію. *STATISTICA* дозволяє досить просто здійснити таку дію. Для цього, наприклад, слід викликати на стовпчику зі змінною, яку потрібно стандартизувати, контекстне меню і вибрати у ньому послідовно пункти **Заполнить/Стандартизовать блок**, **Стандартизовать столбцы (строки)**.

Проведення ієрархічної класифікації

Ієрархічну класифікацію ще називають *таксономією* (від грец. τάξις – строй, порядок і νόμος – закон). Математично таксономією є деревоподібна структура класифікацій певного набора об'єктів. Наверху цієї структури знаходиться єдина класифікація, яка відноситься до усіх об'єктів даної таксономії (вона ще називається *кореневим таксоном*). Таксони, що знаходяться нижче кореневого, є більш специфічними класифікаціями, що відносяться до піднаборів загального набора об'єктів, що підлягають класифікації.

 Перед початком кластеризації за будь-яким методом слід пам'ятати про таке.

1. Потрібно визначитися, над якими даним (у рядках або стовпчиках) слід здійснити кластеризацію. Наприклад, якщо потрібно здійснити кластеризацію річок Полісся, які розташовані по рядках (їх характеристики (змінні) – знаходяться у стовпчиках).

2. Дані мають бути стандартизованими, тобто для кожного даного по стовпцю застосовано дію віднімання середньої величини і ділення одержаного результату на корінь квадратний з дисперсії.

Алгоритм проведення кластеризації має такий вигляд.

1. Виконати команду **Анализ ▸ Многомерный разведочный анализ ▸ Кластерный анализ**. З'явиться вікно вибору метода кластеризації.

2. Вибрати у вікні метод "Иерархическая классификация".

3. Натиснути **«ОК»**. З'явиться вікно настроювань метода.

4. *Визначення об'єктів*. Якщо кластеризації підлягають об'єкти (а вони розташовані у рядках), то зі списку "Объекты" слід вибрати "Наблюдения (строки)".

5. *Вибір характеристик (змінних) для аналізу*. Натиснути кнопку **«Переменные» («Variables»)** і вибрати змінні для аналізу.

6. Зі списку "Мера близости" вибрати міру подібності, наприклад, евклідову метрику, яка пропонується за замовчуванням.

7. Зі списку "Правило объединения" вибрати правило об'єднання.

8. Натиснути **«ОК»**. З'явиться вікно з результатами ієрархічної класифікації.

Найбільш цікавими для дослідника є графіки, які називаються *дендрограмами*. Дендрограми показують близькість значень набора даних по одному з параметрів, використовуючи ось *y* для розстановки самих значений, а ось *x* – величини параметра. Графік відображається у вигляді набора з'єднаних одна з одною горизонтальних ліній, які з'єднуються, якщо значення співпадають по параметру. При цьому чім раніше збіг значень знаходиться по оси *x*, тем ближче вони є одна до одної.

Проведення кластерного аналізу методом К-середніх

Ідея кластеризації ітеративним *методом К-середніх*: об'єкт відносять до того класу, відстань до якого є мінімальною. Відстань розуміється як евклідова, тобто об'єкти розглядаються як точки евклідового простору.

Наприклад, при кластеризації річок Полісся завданням методу є розподіл річок на кілька груп, в яких річки мало відрізняються одна від одної (значно менше, ніж за сукупністю у цілому). Це завдання є складним, оскільки порівняння річок здійснюється за багатьма характеристиками.

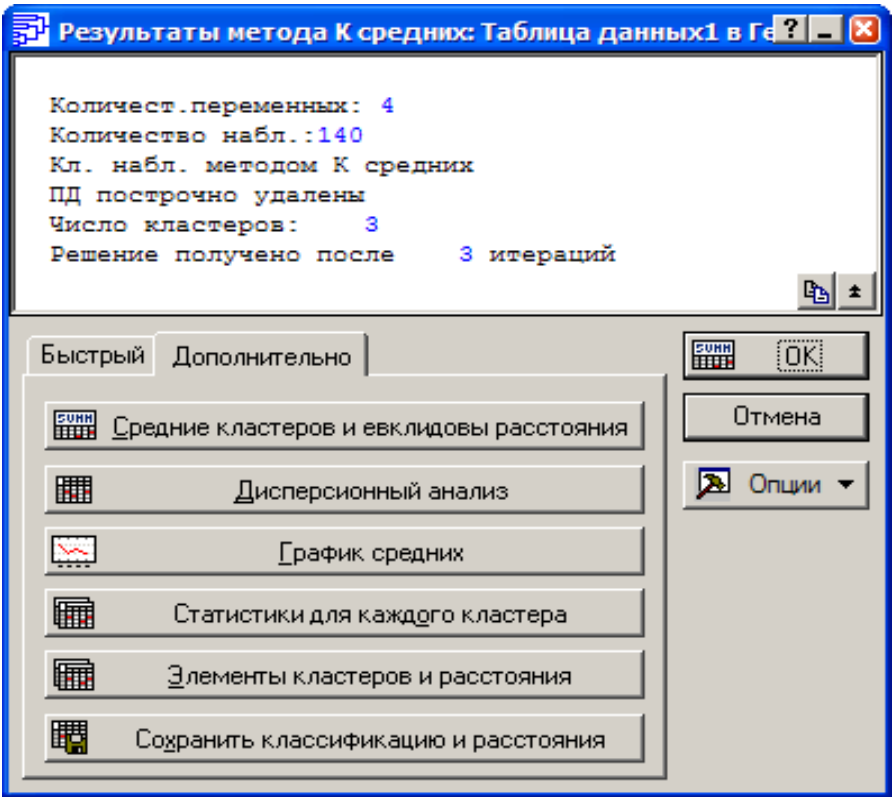
Алгоритм проведення аналізу має такий вигляд.

1. Виконати команду **Анализ ▸ Многомерный разведочный анализ ▸ Кластерный анализ**. З'явиться вікно вибору метода кластеризації.
2. Вибрати у вікні метод "Кластеризация методом К средних".
3. Натиснути **«ОК»**. З'явиться вікно настроювань метода.
4. *Визначення об'єктів*. Якщо кластеризації підлягають об'єкти (а вони розташовані у рядках), то на вкладниці "*Быстрый*" (швидкій) зі списку "Объекты" слід вибрати "Наблюдения (строки)".
5. *Вибір характеристик (змінних) для аналізу*. Натиснути кнопку **«Переменные» («Variables»)** і вибрати змінні для аналізу.
6. Визначити кількість кластерів у полі "Число кластеров".
7. Визначити кількість ітерацій для побудови кластерів у полі "Число итераций".
8. Натиснути **«ОК»**. З'явиться вікно "*Результаты метода К*

средних".

Аналіз результатів

Верхня частина вікна результатів містить зведені дані для аналізу, а також інформацію відносно того, після скількох ітерацій побудовані кластери.



Вкладка "*Быстрый*".

"График средних" відображає у вигляді лінійного графіка середні значення для кожного кластера.

Вкладка "*Дополнительно*".

"Элементы кластеров и расстояний". Цей інструмент дозволяє переглянути згруповані за кластерами об'єкти.

І, нарешті, якщо потрібно зберегти результати кластеризації, то слід вибрати інструмент "Сохранить классификацию и расстояния". Після цього з'явиться вікно, в якому буде запропоновано вибрати потрібні змінні. До цих змінних будуть додані стовпчики з номером спостереження, номер кластеру, до якого належіть об'єкт, а також евклідовою відстанню.  При цьому зручно, якщо один із стовпчиків буде містити *імена спостережень* – назви об'єктів.

Для створення імен спостережень слід виконати такі дії:

1. Викликати контекстне меню на будь-якому з номерів рядків.
2. Вибрати з нього пункт "Диспетчер имен наблюдений…".
3. Переконатися що перемикач у групі "Перенести имена наблюдений" має положення "из".
4. Встановити курсор в поле "Переменная", натиснути клавішу **<F2>** для виклику списку змінних і вибрати з нього змінну, значення якої будуть використовуватися як імена спостережень.
5. Натиснути **«ОК»**.
6. Якщо у вікні не була задана довжина поля, яка визначається за іменами спостережень, то система запропонує самостійно визначити розмірність стовпчика з іменами.

Імовірнісний калькулятор

Імовірнісний калькулятор замінює паперові таблиці ймовірнісних розподілів. За його допомогою можна ефективно вирішувати багато статистичних завдань. На практиці він є засобом, який дозволяє швидко побудувати графіки найбільш розповсюджених функцій розподілу. Алгоритм використання імовірнісного калькулятора такий.

1. Завантажити калькулятор, що можна зробити двома шляхами:

- Завантажити модуль "Основные статистики и таблицы" ("Bases Statistics/Tables"), вибрати у вікні "*Основные статистики*

и таблицы" ("Basic Statistics and Tables:") пункт "Вероятностный калькулятор" ("Probability Calculator", імовірнісний калькулятор) і натиснути **«ОК»**.

- Виконати команду **Анализ ▸ Вероятностный калькулятор ▸ Распределения…**

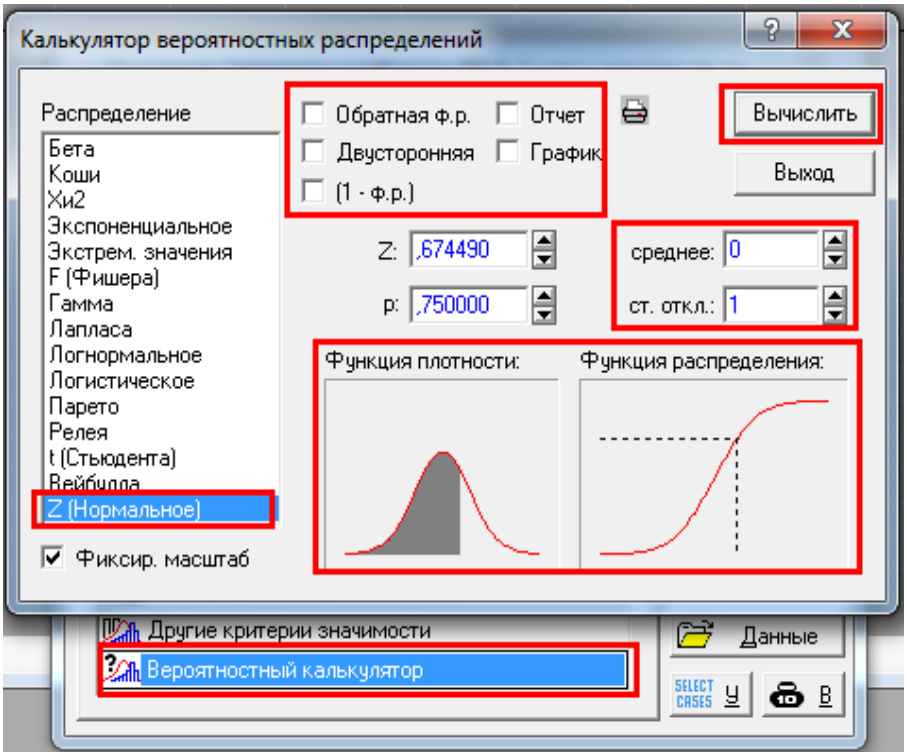
2. З'явиться вікно "*Калькулятор вероятностных распределений*" ("Probability Distribution Calculator", калькулятор імовірнісних розподілів). Список розподілів містить найбільш поширені розподіли: нормальний, логнормальний, розподіл Стьюдента і т. ін., з якого вибирається потрібний розподіл.

3. Після вибору розподілу у вікні з'являються поля-параметри для вибраного розподілу і відображаються графіки щільності та розподілу. При зміні параметрів автоматично змінюються форми цих графіків.

4. У верхній частині вікна розташовано кілька додаткових полів з таким призначенням:

- Обратная ф.р. (Inverse, зворотна функція розподілу).
- Двусторонняя (Two-tailed, двостороння функція).
- Встановлення прапорця для поля-мітки "График" (Create Graph) ініціює при розрахунку побудову графіка щільності і функції розподілу.

5. Після введення значень параметрів функції і натискання кнопки **«Вычислить» («Compute»**, Обчислити) відбувається розрахунок відповідного значення (наприклад, для нормального розподілу – відповідного квантиля).



Нормальний розподіл

Нормальний розподіл *(закон Гауса)* найчастіше зустрічається у математичній статистиці і теорії ймовірностей при вивченні природних явищ. Наприклад відомо, що випадкові похибки в економічних рядах, радіотехніці і т. ін. мають приблизно нормальний розподіл; за його допомогою можна наближено описати зріст дорослих людей тощо. Крива такого типу розподілу описується відповідною функцією, вона є симетричною, її праве крило абсолютно ідентичне лівому. Нормальна крива розподілу зображує статистичну сукупність, в якій відхилення індивідуальних значень ознаки від середньої в один бік зустрічаються так саме часто, як і відхилення в інший бік. Емпірична крива розподілу завжди більшою чи меншою мірою відхиляється від теоретичної кривої.

Нормальний розподіл характеризується двома параметрами: середньою величиною і стандартним відхиленням. За допомогою імовірнісного калькулятора дуже просто побудувати нормальний розподіл за будь-якими значеннями цих параметрів та одержати ймовірнісні оцінки. Разом із тим калькулятор дає змогу виконати і зворотні обчислення: за заданими значеннями параметрів обчислити рівень імовірності.

Приклад. Відомо, що у регіоні зріст дорослих чоловіків наближено має нормальний розподіл із середньою 176 см і стандартним відхиленням 7,63 см. Яка імовірність того, що зріст випадково вибраного чоловіка не більше 183 і не менше 173 см?

Розв'язок.

1. Зі списку розподілів вибрати "Z Нормальное [распределение]" ("Z Normal").
2. Встановити для середньої величини значення 176 см і для стандартного відхилення 7,63 см.
3. У полі "Z" ввести значення "183" і обчислити значення "p".
4. Виконати аналогічні дії для значення 173 см.
5. Різниця між першим і другим значення і буде імовірністю того, що зріст випадково вибраного чоловіка знаходиться в межах від 173 см до 183 см.

Розподіл хі-квадрат (χ^2)

Випадкова величина, що має розподіл χ^2 визначається як сума квадратів *k* незалежних стандартних нормальних величин (тобто величин, які мають нормальний стандартний розподіл). Число *k* для χ^2 називається числом ступенів свободи. Якщо *k* дорівнює "1", то випадкова величина χ^2 дорівнює квадрату стандартної нормальної величини. Отже, розподіл χ^2 характеризується тільки одним параметром – числом ступенів свободи.

χ^2 використовується, наприклад, під час перевірки залежностей у таблицях спряженості. Для цього розраховується теоретичне

значення квантиль χ^2-розподілу для певної кількості ступенів свободи та обраного рівня ймовірності і порівнюється з фактичним. Якщо фактичне значення більше за його теоретичне значення, то це є ознакою зв'язку між досліджуваними змінними Для побудови χ^2-розподілу слід виконати такі дії.

1. Зі списку розподілів вибрати "Хи2" ("Chi").
2. У полі "ст. св." ("df") ввести число ступенів свободи, в полі "p" – імовірність.
3. Натиснути кнопку **«Вычислить» («Compute»)**, після чого в полі "Хи2" ("Chi") з'явиться квантиль χ^2-розподілу для заданого значення числа ступенів свободи.

t-розподіл Стьюдента

t-розподіл є важливим у тих випадках, коли розглядаються оцінки середньої з невідомою загальною дисперсією вибірки. У цьому разі використовують вибіркову дисперсію і t-розподіл. t-розподіл з'являється у таблицях виведення регресійного аналізу, при аналізі динамічних рядів. Поряд з нормальним розподілом і розподілом χ^2 він є одним з найважливіших розподілів.

Для побудови t-розподілу слід виконати такі дії.

1. Вибрати зі списку розподілів "t (Стьюдента)" ("t (Student))".
2. У полі "ст. св." ("df") ввести число ступенів свободи, в полі "p" – імовірність.
3. Натиснути кнопку **«Вычислить» («Compute»)**, після чого в полі "t" з'явиться квантиль χ^2-розподілу для введеного значення числа ступенів свободи.

 За великих значень числа ступенів свободи (більше 30) *t*-розподіл практично співпадає із стандартним нормальним розподілом.

Лабораторні роботи

Загальні вказівки до виконання:

1. "*Індивідуальний номер"* – Ваш номер у навчальному журналі групи (курсу).
2. Всі роботи зберігаються у власній папці.

Лабораторна робота №1

Тема. Знайомство з програмою. Введення даних

Мета. Навчитися створювати файл даних, вводити дані спостереження, а також здійснювати їх первинний аналіз.

Вказівки до виконання:

Одержати список змінних у вікні додавання нової змінної можна здійснивши подвійне клацання мишею в полі "После" або натиснути клавішу **<F2>**.

Завдання

1. Завантажте STATISTICA.
2. Створіть файл даних (електронну таблицю) для 2-х змінних та 5-ті спостережень.
3. Створіть змінні "Чисельність персоналу" та "Обсяг продукції, млн. грн." використовуючи дані таблиці 4. Тип даних – цілий (Integer). Зверніть увагу на ім'я змінної, яке створює система за замовчуванням.  Якщо під час введення даних виникнуть проблеми, то з'ясуйте причину її появи (проаналізуйте типи даних) і усуньте помилку.
4. Додайте після змінної "Обсяг продукції, млн. грн." змінну "Обсяг продукції в розрахунку на одного працюючого, грн.".
5. Створіть формулу для автоматичного обчислення змінної "Обсяг продукції в розрахунку на одного працюючого, грн." за існуючими змінними.
6. Відобразіть у заголовку даних імена змінних разом:
   - з типами даних;
   - з формулами.

7. Відобразіть таблицю змінних з усіма їх властивостями.

8. Уведіть загальну інформацію про таблицю з даними: "Дані про обсяг продукції і чисельність персоналу", а потім Ваше прізвище і групу. Застосуйте до тексту курсив і напівжирне накреслення.

9. Виконайте настроювання системи для того, щоб дані у клітинках з формулами при зміні даних, на підставі яких вони обчислюються, не розраховувалися автоматично, тобто змінювалися тільки в ручному режимі.

10. Введіть в електронну таблицю для змінних "Чисельність персоналу" та "Обсяг продукції, млн. грн." частину даних з таблиці 1 і переконайтеся, що значення змінної "Обсяг продукції в розрахунку на одного працюючого, грн." не розраховуються.

11. Розрахуйте дані змінної "Обсяг продукції в розрахунку на одного працюючого, грн." вручну.

12. Виконайте настроювання системи для того, щоб дані у клітинках з формулами при зміні даних, на підставі яких вони обчислюються, розраховувалися автоматично.

13. Введіть в електронну таблицю для змінних "Чисельність персоналу" та "Обсяг продукції, млн. грн." дані з таблиці 4, що залишилися, і переконайтеся, що значення змінної "Обсяг продукції в розрахунку на одного працюючого, грн." під час введення даних розраховуються автоматично.

14. Впорядкуйте значення змінної "Чисельність персоналу" за спаданням, а потім за зростанням. Для повернення до початкового розташування даних достатньо здійснити відміну здійсненної дії (сортування).

15. Здійсніть настроювання системи для автоматичного збереження даних, інтервал кількості хвилин для збереження задайте на власний розсуд.

16. Збережіть таблицю з даними з довільним іменем (але надалі у лабораторних роботах вона буде називатися "Обсяг продукції").

17. Закрийте таблицю з даними.

18. Створіть нову таблицю з даними.

19. Створіть змінну "Заробітна плата робітників". Тип даних – байтовий (Byte).

20. За наведеними нижче даними визначте потрібну кількість спостережень і додайте їх в таблицю з даними.

21. Введіть дані щодо заробітної плати: 3300, 2600, 3100, 2000, 2400, 3200, 2500, 2200, 2600, 2500, 2600, 2300, 2600, 3200, 2400, 3400, 1400, 2300, 2200, 3100, 3000, 3000, 2300, 1800, 40000.

 Якщо під час введення даних виникнуть проблеми, то з'ясуйте причину її появи (проаналізуйте типи даних) і усуньте помилку.

22. У модулі перевірки даних (**Данные ▸ Проверка данных** (**Data ▸ Verify Data**)) здійсніть аналіз даних з метою виявлення невірних даних. Визначте з цією метою дві умови, згідно яким значення змінної мають знаходиться в інтервалі від 1600 до 4000 включно.

23. Якщо під час аналізу такі значення будуть знайдені, то замініть їх на найближчі до них значення умови.

24. Збережіть таблицю з даними:

- у вигляді робочої книги STATISTICA (ім'я таблиці – довільне, але надалі у лабораторних роботах вона буде називатися "Заробітна плата");
- у вигляді Web-сторінки;
- як файл електронної таблиці Excel.

25. Додайте змінну "Категорія робітників". У вікні властивостей змінної задайте такі її властивості:

- Тип даних – байтовий.
- Натисніть кнопку **«Текстовые метки»** ("Text Labels", текстові мітки) і у вікні редактора текстових меток створіть три текстові метки (значення у полі "Число" (Numeric) залишайте таким, що його створює система автоматично, наприклад, 101):

25.1. Керівники.

25.2. Службовці.

25.3. Робітники.

26. Введіть дані для змінної "Категорія робітників": 101, 102, 103, 103, 102, 102, 103, 102, 103, 103, 102, 103, 103, 102, 102, 101, 103, 103, 102, 102, 102, 102, 103, 103, 101. Загальна кількість даних

має співпадати із загальною кількістю даних змінної "Заробітна плата робітників".

27. Перемкніть відображення даних з числових значень на текстові і навпаки.

28. Збережіть таблицю з даними.

Таблиця 4

Початкові дані до теми "Введення даних"

Чисельність персоналу	Обсяг продукції, млн. грн.
890	200,05
845	188, 18
1130	307, 77
925	224, 24
700	112, 11

Контроль знань та навичок

Після виконання лабораторної роботи студент повинен ***знати*:**

1. Як називають файл з даними STATISTICA?
2. Що являє собою "спостереження" у таблиці з даними?
3. Що являє собою "змінна" у таблиці з даними?
4. Як створити нову таблицю з даними?
5. Яке ім’я за замовчуванням надає система новій змінній?
6. Чи можна змінити розмірність вже створеної таблиці із даними, додаючи або вилучаючи спостереження та змінні?
7. Як вилучити змінну?
8. Як додати нову змінну у конкретному місці?
9. Як додати або вилучити спостереження?
10. Що таке "властивості" змінної?
11. Які типи даних є у STATISTICA?
12. У чому відмінність між числовими типами даних? Наведіть приклади.
13. Які можна ввести не числові дані?
14. Що таке "Текстові метки"?
15. Як відобразити таблицю змінних з усіма їх властивостями?

16. Яким чином можна відобразити у клітинці з ім’ям змінної тип даних і розрахункові формули?

17. Для чого призначена загальна інформація про таблицю з даними?

18. Як ввести загальну інформацію про таблицю з даними?

19. Коли доцільно подати змінну у вигляді формули? Наведіть приклад.

20. З яких елементів складається розрахункова формула для змінної?

21. Як для змінної створити розрахункову формулу?

22. Як відобразити не числові (текстові) дані у текстовому вигляді або вигляді їх числових еквівалентів?

23. Які дії слід застосувати для того, щоб дані в клітинках із формулами обчислювалися автоматично?

24. Які дії слід застосувати для того, щоб дані в клітинках із формулами обчислювалися в ручному режимі?

25. Як здійснити сортування (упорядкувати) значення змінної?

26. Як здійснити аналіз даних, накладаючи для цього умови на значення даних?

27. У яких форматах можна зберігати файл з даними STATISTICA?

Після виконання лабораторної роботи студент повинен ***уміти:***

1. Створити нову таблицю з даними.
2. Змінити розмірність таблиці із даними.
3. Вилучити або додати нову змінну у конкретному місці.
4. Додати або вилучити спостереження.
5. Відобразити таблицю змінних з усіма їх властивостями.
6. Відобразити у клітинці з ім’ям змінної тип даних і розрахункові формули.
7. Для чого призначена загальна інформація про таблицю з даними?
8. Як ввести загальну інформацію про таблицю з даними?
9. Створити для змінної розрахункову формулу.
10. Ввести нечислові дані; створювати для них текстові мітки.

11. Відобразити нечислові (текстові) дані у текстовому вигляді або вигляді їх числових еквівалентів.

12. Здійснити настроювання системи, щоб дані в клітинках із формулами обчислювалися автоматично або в ручному режимі.

13. Відсортувати (упорядкувати) значення змінної.

14. Здійснити аналіз даних, накладаючи для цього умови на значення даних.

15. Зберегти файл з даними у вигляді Web-сторінки, як файл електронної таблиці Excel.

Лабораторна робота №2

Тема. Частотний аналіз. Побудова графіків

Мета. З'ясувати принципи групування даних і застосування їх для аналізу даних.

Завдання

1. Завантажте таблицю з даним "Заробітна плата".

2. Визначте мінімальне та максимальне значення для змінної, здійснив впорядкування (сортування) змінної.

3. За одержаними значеннями вручну розрахуйте кількість груп за формулою Стерджеса.

4. Вручну розрахуйте ширину (рівного) інтервалу.

5. Зверніться до модуля "Частотный анализ" ("Frequency tables") з меню "Анализ" (Statistics) різними шляхами (після появи вікна "Частотный анализ", "Frequency tables" у перших двох випадках просто закривайте його):

- через команду основного меню;
- використовуючи панель інструментів "Анализ" (Statistics) (за відсутності панелі виведіть її використовуючи стандартну команду з пункту головного меню Вид (View);
- застосуванням кнопки  "Вызвать меню часто используемых средств" ("Start menu", "Викликати меню засобів, що найчастіше використовуються").

6. Відкрийте вікно модуля "Частотный анализ" ("Frequency

tables") з меню "Анализ" (Statistics). Виконайте у модулі наведені нижче дії. Всі одержані результати під час виконання цих дій зберігайте в *одній* робочій книзі, за закінченням виконання усіх дій збережіть на диску саму робочу книгу. Виконайте такі дії:

- Побудуйте таблицю частот і гістограму не змінюючи параметрів модуля. З'ясуйте, які саме показники (окрім частот) розраховуються системою автоматично за замовчуванням. Перейдіть на вкладку "Опции" (Options) і перегляньте, які ще статистики можна розрахувати в модулі.
- Побудуйте таблицю частот і гістограму за розрахованою Вами кількістю груп. Проаналізуйте відмінності цих даних від попередніх.
- Побудуйте таблицю частот і гістограму за розрахованим Вами кроком інтервалу; як початкове значення першого інтервалу виберіть мінімальне з усіх значень.
- Перевірте розподіл на нормальність за критерієм Шапіро-Уїлка. Для цього перейдіть на вкладку "Нормальность" (Normality), встановить позначку для поля "Критерий Шапиро-Уилка" ("Shapiro-Wilk's W test") і натисніть кнопку **«Тест на нормальность»** (**«Test for normality»**).

7. Зверніться до модуля "2M Гистограмма" ("2D Histograms") з меню "Графика" (Graphs). Виконайте у модулі наведені нижче дії. Всі створені діаграми під час виконання цих дій зберігайте в *одній* робочій книзі, за закінченням виконання усіх дій збережіть на диску саму робочу книгу. Виконайте такі дії:

- Побудуйте гістограму не змінюючи параметрів модуля. Визначте, які спостереження система включила до кожного інтервалу (для цього достатньо навести курсор на стовпчик гістограми).
- Побудуйте гістограму за розрахованою Вами кількістю груп.
- Відобразіть на діаграмі відсотки. Для цього на вкладці "Дополнительно" ("Advanced") встановіть прапорець для поля-мітки "Показать проценты" ("Show percentages").
- Перевірте розподіл на нормальність за критерієм Шапіро-

Уїлка.

- Побудуйте графічно кумуляту розподілу (Накопичена частота має назву *кумулятивної* або *кумуляти*). Для цього на вкладці "Дополнительно" ("Advanced") зі списку "Отображаемый тип" ("Showing Type") виберіть "Кумулятивный" ("Cumulative").
- Побудуйте графічно огіву розподілу (Огіва є дзеркальним відображенням кумуляти. При її побудові на осі абсцис відкладають накопичені частоти або частки, а на осі ординат – межі інтервалів варіаційного ряду). Для цього на вкладці "Параметры 2" ("Options 2") зі списку "Позиции осей X-У" ("X-Y Axes position") виберіть "Обратный" ("Reversed").

8. З'ясуйте, в якому режимі відображається графік. Перемкніть відображення в інший режим.

9. Перейдіть до режиму редагування графіку і виконайте такі дії:

- змініть загальну назву графіку на "Гістограма для змінної "Заробітна плата робітників";
- знайдіть місце, де можна змінити максимальне та мінімальне значення на осі значень і встановіть мінімальне значення "1660". Поясніть зміну вигляду графіка.

10. Додайте у таблицю з даними після змінної "Категорія робітників" ще дві змінні: "Значення заробітної плати", "Частоти". Тип даних – числовий.

11. Застосуйте модуль "Основные статистики и таблицы" ("Bases Statistics/Tables") (пункт "Таблицы частот", "Frequency tables") для побудови таблиці частот змінної "Заробітна плата робітників". Таблицю будуйте таким чином, щоб для кожного значення змінної створювалася окрема група.

12. Скопіюйте з таблиці частот стовпчик "Категории" (Category), що містить значення заробітної плати і вставте їх у змінну "Значення заробітної плати".  Не звертайте увагу, що кількість спостережень для цієї змінної буде менше, ніж кількість спостережень для інших змінних.

13. Скопіюйте з таблиці частот стовпчик "Частота" (Count), що містить розраховані значення часто і вставте їх у змінну "Частоти".

14. Використовуючи дані змінних "Значення заробітної плати"

та "Частоти" побудуйте полігон розподілу. Для цього виконайте такі дії:

- Виконайте команду **Графики ▸ 2M Графики ▸ Линейные графики (по переменным)** (**Graphs ▸ 2D Graphs ▸ Line Plots (Variables)**), що спричиніть появу вікна "2M Линейные графики" (2D Line Plot).
- На вкладці "Дополнительно" (Advanced):

14.1. У групі "Тип графика" (Graph Type) виберіть "XY Trace".

14.2. Натисніть кнопку «**Переменные**» («**Variables**») і визначить значення по осі X ("Значення заробітної плати") і Y ("Частоти").

14.3. Зі списку "Отображать точки" (Display point) виберіть "On".

15. Перетягніть зі списку вікна робочої книги назву графіку за межі вікна. Збережіть нове вікно з графіком у якомусь графічному форматі. Для виклику дій, що можна здійснювати з вмістом вікна викликайте в його площині контекстне меню.

16. Збережіть вікно з графіком:

- у графічному форматі STATISTICA;
- у форматі GIF;
- у форматі BMP.

17. Збережіть файл даних.

Контроль знань та навичок

Після виконання лабораторної роботи студент повинен ***знати*:**

1. Як можна звернутися до конкретного модуля?
2. Для чого призначено модуль "Частотный анализ" ("Frequency tables")?
3. Як називають вікно виведення результатів у STATISTICA?
4. Яка структура вікна виведення результатів?
5. Як побудувати таблицю частот?
6. Як задати ширину (крок) інтервалу для побудови групування?
7. Які можна задати початкові значення для першого інтервалу групування?
8. Як задати кількість груп для побудови групування?

9. Які показники можна одержати у модулі "Частотный анализ" ("Frequency tables")?

10. Яким варіанти побудови групувань надає користувачу модуль "Частотный анализ" ("Frequency tables")?

11. Як здійснити перевірку розподілу на нормальність за допомогою критерію Шапіро-Уїлка у модулі "Частотный анализ" ("Frequency tables")?

12. Як побудувати гістограму у модулі "Частотный анализ" ("Frequency tables")?

13. Які способи побудови гістограми має програма?

14. Як побудувати гістограму у модулі "2М Гистограмма" ("2D Histograms")?

15. Як задати кількість інтервалів групування у модулі "2М Гистограмма" ("2D Histograms")?

16. Як у модулі "2М Гистограмма" ("2D Histograms") визначити спостереження, що увійшли до конкретного інтервалу діаграми?

17. Як у модулі "2М Гистограмма" ("2D Histograms") відобразити відсотки на діаграмі?

18. Як здійснити перевірку розподілу на нормальність за допомогою критерію Шапіро-Уїлка у модулі "2М Гистограмма" ("2D Histograms")?

19. Що таке кумулята розподілу?

20. Як побудувати графічно кумуляту розподілу у модулі "2М Гистограмма" ("2D Histograms")?

21. Що таке огіва?

22. Як побудувати огіву у модулі "2М Гистограмма" ("2D Histograms")?

23. Які є режими відображення графіка?

24. В якому режимі за замовчуванням відображається графік?

25. Як в режимі редагування графіка змінити загальну назву графіка, назви його осей?

26. Як в режимі редагування графіка змінити максимальне та мінімальне значення на осі абсцис?

27. Які результати аналізу можна зберігати в одній робочій книзі?

28. Як у файл з даними додати нову змінну?
29. У яких форматах можна зберегти графік?

Після виконання лабораторної роботи студент повинен ***уміти:***

1. Звернутися до конкретного модуля різними способами.
2. Побудувати таблицю частот.
3. Змінити ширину (крок) інтервалу для побудови групування.
4. Визначити початкові значення для першого інтервалу групування.
5. Задати кількість груп для побудови групування.
6. Побудувати групування різними способами у модулі "Частотный анализ" ("Frequency tables").
7. Здійснити перевірку розподілу на нормальність за допомогою критерію Шапіро-Уїлка у модулі "Частотный анализ" ("Frequency tables").
8. Побудувати гістограму.
9. Задати кількість інтервалів групування у модулі "2М Гистограмма" ("2D Histograms").
10. Визначити спостереження, що увійшли до конкретного інтервалу діаграми.
11. Відобразити відсотки на діаграмі.
12. Здійснити перевірку розподілу на нормальність за допомогою критерію Шапіро-Уїлка у модулі "2М Гистограмма" ("2D Histograms").
13. Побудувати графічно кумуляту розподілу.
14. Побудувати огіву.
15. Змінити режими відображення графіка.
16. Здійснити редагування графіка, зокрема змінити загальну назву графіка, назви його осей.
17. У режимі редагування графіка змінити максимальне та мінімальне значення на осі абсцис.
18. Додати нову змінну у файл з даними.
19. Зберігати результати аналізу в одній робочій книзі.
20. Зберегти графік у різних форматах.

Лабораторна робота №3

Тема. Описові статистики

Мета. З'ясувати принципи розрахунку описових статистик та використання їх для аналізу даних.

Вказівки до виконання:

1. Визначення переліку додаткових статистик, що потрібно розрахувати, в модулі "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") здійснюється на вкладці "Дополнительно" ("Advanced").

2. Англомовні скорочення на діаграмі розмаху:

- SE – Standard error of estimate Standard (Стандартна похибка оцінювання, Стандартная ошибка оценки).
- SD – Standard Deviation (Стандартне відхилення, Стандартное отклонение).

Завдання

1. Завантажте таблицю з даними "Обсяг продукції".
2. Застосуйте блокові статистики для змінних "Чисельність персоналу" та "Обсяг продукції, млн. грн." з метою визначення:
   - середніх величин;
   - всіх описових статистик.
3. Завантажте таблицю з даними "Заробітна плата".
4. Застосуйте блокові статистики з метою визначення середньої величини для перших десяти спостережень.
5. Застосуйте модуль "Гистограммы" ("Histograms") з меню "Графика" (Graphs) для розрахунку базових описових статистик. З'ясуйте, які саме статистики при цьому автоматично розраховуються. Для розрахунку статистик встановить позначку для полямітки "Описательные статистики" ("Descriptive Statistics") на вкладці "Дополнительно" ("Advanced") в групі "Статистики" ("Statistics").
6. Застосуйте модуль "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") для розрахунку базових описових статистик, що розраховуються за замовчуванням. З'ясуйте, що це за статистики.

 Виконаний аналіз і всі аналізи, що будуть здійснені надалі, запам'ятовуйте в одній робочій книзі.

7. Виконайте в цьому модулі такі дії:

- Розрахуйте моду (Mode) та медіану (Median). Порівняйте значення середньої величини, моди та медіани. Вмійте пояснити їх сутність. Проаналізуйте результати аналізу і визначте, що являє собою показник "Частота моды" (Frequency of Mode).
- Розрахуйте перший (нижній) та останній (верхній) квартилі (Lower, Upper Quartile), перший та останній децилі (процентилі) (Percentile). Дайте пояснення сутності цих показників.
- Розрахуйте показники варіації: "Размах вариации" (Range), дисперсію (Variance), "Стандартное отклонение" (Standard Devation, стандартне відхилення), і "Стандартная ошибка (середнього значення)" (Std. err. of mean, стандартна похибка середнього значення). Дайте пояснення сутності показників розмах варіації та дисперсія.
- Розрахуйте характеристики форми розподілу "Асимметрия" (Skewness) і "Эксцесс" (Kurtosis). На їх підставі зробіть висновки щодо напряму асиметрії і міри скошеності; вмійте обґрунтувати свої висновки.
- Побудуйте гістограму, під час її побудови задайте для неї розраховану Вами кількість інтервалів. Нанесіть на гістограму напис з Вашим висновком відносно напряму асиметрії і міри скошеності.
- Побудуйте діаграму розмаху, для чого слід натиснути кнопку «**Диаграмма размаха для всех переменных» («Box & Whisker plot for all variables»**). Визначте на вкладці "Опции" (Options), які опції для діаграми можна застосовувати в модулі.

8. Побудуйте діаграми розмаху в модулі "*2M Диаграммы размаха*" ("2D Box Plots"). Зберігайте в одній робочій книзі всі побудовані надалі діаграми.

- У вікні настройок модуля здійсніть таке:
  - 8.1. Визначте, яка змінна буде змінною групування для змінної "Заробітна плата робітників". Задайте ці змінні.
  - 8.2. З'ясуйте, який саме показник центру розподілу відображатиметься на графіку, а також форму (стиль) його відображення.
  - 8.3. Побудуйте діаграму.
  - 8.4. Здійсніть аналіз діаграми:

8.4.1. Вмійте поясніти, що відображає кожний з її елементів.

8.4.2. Вмійте поясніти, які саме настройки і як формують вигляд діаграми: її вуси, основу коробки, наявність викидів.

8.4.3. Поясніть, чому одна з діаграм не має вусів.

8.4.4. З'ясуйте, як можна визначити, які спостереження складають тут чи іншу групу.

8.5. Дайте відповідь на питання: чим відрізняються можливості побудови діаграми розмаху в модулі від тих, що має користувач в модуль "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables")?

- Змініть стиль відображення центру розподілу на риску. Побудуйте діаграму. Здійсніть її аналіз.
- Побудуйте діаграму, вибравши як показник центру розподілу середню арифметичну. Поясніть наявність на "вусах" окремих діаграм точок-викидів.
- Побудуйте таку саме діаграму, але без викидів.
- Побудуйте діаграму для категорій "Службовці" і "Робітники".

9. З'ясуйте, які саме графічні елементи повинна мати діаграма розмаху для змінної "Заробітна плата робітників". Обґрунтуйте свій висновок.

 Контроль знань та навичок

Після виконання лабораторної роботи студент повинен ***знати*:**

1. Що таке "описові статистики"? Назвіть найбільш відомі з них.
2. Що таке "блокові статистики"?
3. Які описові статистики можна розрахувати використовуючи блокові статистики?
4. Як розрахувати описові статистики, використовуючи блокові статистики, в тому числі для групи спостережень?
5. В яких модулях можна розрахувати описові статистики?
6. Який модуль аналізу спеціально призначений для розрахунку описових статистик?

7. Що у статистиці відносять до показників центру розподілу?

8. Як в модулі "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") розрахувати показники центру розподілу?

9. Яка сутність показника "Мода"?

10. Яка сутність показника "Медіана"?

11. Що собою являють перший (нижній) та останній (верхній) квартилі?

12. Як в модулі "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") розрахувати перший (нижній) та останній (верхній) квартилі?

13. Що собою являють децилі (процентилі)?

14. Як в модулі "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") розрахувати децилі (процентилі)?

15. Що у статистиці відносять до показників варіації?

16. Як в модулі "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") розрахувати показники варіації?

17. Яка сутність показника "Розмах варіації"?

18. Яка сутність показника "Стандартне відхилення"?

19. Яка сутність показника "Стандартна похибка (середнього значення)"?

20. Які Ви знаєте характеристики форми розподілу?

21. Як на підставі характеристики форми розподілу зробить висновки щодо напряму асиметрії і міри скошеності розподілу?

22. Для чого призначена діаграма розмаху?

23. З яких графічних елементів складається діаграма розмаху?

24. На підставі чого визначається перелік графічних елементів діаграми розмаху?

25. Як в модулі "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") діаграму розмаху?

26. Як побудувати діаграму розмаху в модулі "*2М Диаграммы размаха*" ("2D Box Plots")?

27. У чому основна відмінність діаграми розмаху між такою, що будується в модулі "Описательные статистики" (Descriptive Statistics) і такою, що будується в модулі "*2М Диаграммы размаха*" ("2D Box Plots")?

28. Що таке "незалежна змінна" і змінна групування для неї?

29. Які настройки в модулі "*2М Диаграммы размаха*" ("2D Box Plots") формують вигляд діаграми розмаху?

30. Як можна визначити, які спостереження складають тут чи іншу групу діаграми розмаху?

31. Як можна змінити стиль відображення центру розподілу на діаграмі розмаху?

32. Що являють собою "викиди" на діаграмах розмаху?

33. Як можна відключити відображення викидів на діаграмах розмаху?

34. Як побудувати діаграму розмаху тільки для окремих груп змінної групування?

Після виконання лабораторної роботи студент повинен ***уміти:***

1. Розрахувати "блокові статистики".

2. Розрахувати описові статистики за допомогою блокових статистик, у тому числі для групи спостережень.

3. Використати модуль "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") для розрахунку показників центру розподілу.

4. Використати модуль "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") для розрахунку квартилів.

5. Використати модуль "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") для розрахунку децилів (процентилів).

6. Використати модуль "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") для розрахунку показників варіації.

7. На підставі характеристики форми розподілу зробить висновки щодо напряму асиметрії і міри скошеності розподілу.

8. Визначити перелік графічних елементів діаграми розмаху.

9. Використати модуль "Описательные статистики" (Descriptive Statistics) аналізу "Основные статистики и таблицы" ("Bases Statistics/Tables") для побудови діаграму розмаху.

10. Побудувати діаграму розмаху в модулі "*2M Диаграммы размаха*" ("2D Box Plots").

11. Визначити, які спостереження складають тут чи іншу групу діаграми розмаху.

12. Змінити стиль відображення центру розподілу на діаграмі розмаху.

13. Відключити відображення викидів на діаграмах розмаху.

14. Побудувати діаграму розмаху для окремих груп змінної групування.

Лабораторна робота №4

Тема. Робота з імовірнісним калькулятором

Мета. Навчитися працювати з імовірнісним калькулятором.

Вказівки до виконання:

1. Рівень імовірності для нормального розподілу задається у полі "p".

2. χ^2-розподіл використовується, наприклад, під час перевірки залежностей у таблицях спряженості. Для цього розраховується теоретичне значення квантиль χ^2-розподілу для певної кількості ступенів свободи та обраного рівня ймовірності і порівнюється з фактичним. Якщо фактичне значення χ^2 більше за його теоретичне значення, то це є ознакою наявності зв'язку між досліджуваними змінними.

Завдання

1. Завантажте будь-який створений раніше файл з даними.
2. Завантажте імовірнісний калькулятор.
3. Виберіть нормальний розподіл і виконайте для нього такі дії:

- Визначте, як змінюються графіки функцій зі зміною величини ймовірності за фіксованих значеннях інших параметрів.
- Встановіть для розподілу рівень імовірності, який розрахуйте за формулою "Індивідуальний номер": "кількість студентів у групі".

- Визначте, як змінюються графіки функцій зі зміною середньої величини за фіксованого значення стандартного відхилення.
- Визначте, як змінюються графіки функцій зі зміною стандартного відхилення за фіксованого значення середньої.
- Збережіть один із графіків.

4. *Використовуючи нормальний* розподіл, розв'яжіть таке завдання. Відомо, що середній зріст курсантів військового училища становить 165+("Індивідуальний номер"*0,5) см для стандартного відхилення 8,5 см. Визначити імовірність того, що зріст випадково відібраного курсанта знаходиться в межах від 183 см до 180 см.

5. Виберіть розподіл χ^2 (Хи-2).

- Встановіть число ступенів свободи як значення, що відповідає Вашому "індивідуальному номеру"; якщо він більше "20", то взяти значення "20"; рівень імовірності 0,95.
- Розрахуйте квантиль χ^2–розподілу і порівняйте його з табличним з додатку 1. Чи співпадають вони?
- Фактичне значення χ^2 становить 18,0. Чи буде у цьому разі визначена наявність зв'язку між змінними у таблиці спряженості?

Контроль знань та навичок

Після виконання лабораторної роботи студент повинен ***знати*:**

1. Яке призначення ймовірнісного калькулятора?
2. Як звернутися до імовірнісного калькулятора?
3. Якими параметрами характеризується нормальний розподіл?
4. Як змінюються графіки функцій нормального розподілу при зміні середньої величини?
5. Як змінюються графіки функцій нормального розподілу зі зміною стандартного відхилення?
6. Як зберегти графіки функцій?
7. Які параметри нормального розподілу слід задати для того, щоб розрахувати рівень імовірності?
8. Як розрахувати рівень імовірності для нормального розподілу в межах двох значень середньої величини?

9. Для чого призначений квантиль χ^2–розподілу?

10. Які параметри χ^2–розподілу необхідно задати для розрахунку його квантиля?

11. Як за допомогою значення χ^2–розподілу встановити наявність зв'язку між змінними у таблицях спряженості?

Після виконання лабораторної роботи студент повинен ***уміти:***

1. Обчислити рівень імовірності за заданими параметрами нормального розподілу.

2. Розрахувати рівень імовірності для нормального розподілу в межах двох значень середньої величини.

3. Зберегти графіки функцій.

4. Обчислити значення квантиля χ^2 за заданими параметрами.

5. За допомогою значення квантиля χ^2–розподілу встановити наявність або відсутність зв'язку між змінними у таблицях спряженості.

Література

1. Боровиков В.П., Ивченко И.Г. Прогнозирование в системе Statistica в среде Windows. Основы теории и интенсивная практика на компьютере: учебное пособие. – М.: Финансы и статистика, 2006. – 368 с.

2. Боровиков В.П. Популярное введение в программу STATISTICA. – М.: Финансы и статистика, 2000. – 269 с.

3. Фетісов В. С. Математичні та статистичні пакети. – Ніжин: Видавець ПП Лисенко М.М., 2011. – 324 с.

4. Фетісов В.С. Прикладні пакети статистичної обробки: лабораторний практикум – Ніжин: Видавництво НДУ ім. М.Гоголя, 2010. – 27 с.

Додатки

Додаток 1

Квантилі χ^2 -розподілу при $\alpha = 0{,}05$

k	**0,95**	***k***	**0,95**
1	3,84	11	19,68
2	5,99	12	21,03
3	7,82	13	22,36
4	9,49	14	23,69
5	11,07	15	25,00
6	12,59	16	26,30
7	14,07	17	27,59
8	15,51	18	28,87
9	16,92	19	30,14
10	18,31	20	31,41

Навчальне видання

Фетісов Валерій Сергійович

ПАКЕТ СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ
STATISTICA

Навчальний посібник

Технічний редактор – І. П. Борис

Книга друкується за макетом замовника

Підписано до друку	Формат 60x84/16	Папір офсетний
Гарнітура Times New Roman	Обл.-вид. арк. 3,25	Електр. вид.
Замовлення №	Ум. друк. арк. 6,04	

ВИДАВНИЦТВО НДУ ім.М.В.ГОГОЛЯ

Ніжинський державний університет
імені Миколи Гоголя.
м. Ніжин, вул. Воздвиженська, 3/4
(04631)7-19-72
E-mail: vidavn_ndu@mail.ru
www.ndu.edu.ua

Свідоцтво суб'єкта видавничої справи
ДК № 2137 від 29.03.05 р.