Лекція №4.

ТЕМА 4 ВИКОРИСТАННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ РЕАЛІЗАЦІЇ ЕМПІРИЧНИХ МЕТОДІВ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

4.1 Використання електронних таблиць Exel

При проведенні складного статистичного або інженерного аналізу можна спростити процес і заощадити час, використовуючи крім статистичних функцій надбудову "**Пакет аналізу**" **Exel 2010 [16]**. Для аналізу даних за допомогою цього пакету слід вказати вхідні дані і вибрати параметри; розрахунок буде виконаний за допомогою відповідної статистичної або інженерної макрофункції, а результат буде поміщений у вихідний діапазон. Деякі інструменти дозволяють представити результати аналізу в графічному виді.

Нижче описані інструменти, включені в пакет аналізу. Для доступу до цих інструментів натисніть кнопку **Аналіз даних** в групі **Аналіз** на вкладці **Дані**. Якщо кнопка **Аналіз даних** недоступна, необхідно завантажити надбудову "Пакет аналізу".

Завантаження пакету аналізу:

1. На вкладці Файл выберите команду Параметри, а потім - категорію Надбудови.
2. У списку Управление выберите пункт Надбудови Excel и натисніть кнопку Перейти.
3. У вікні Доступні надстройкиустановите прапорець Пакет анализаи натисніть кнопку ОК.

Порада: Якщо пункт Пакет анализаотсутствует в списку Доступні надбудови, натисніть кнопку Огляд, щоб знайти надбудову.

Якщо виводиться повідомлення про те, що надбудова "Пакет аналізу" не встановлена на комп'ютері, натисніть кнопку Дадля її установки.

Примітка. Для включення в "Пакет аналізу" функцій Visual Basic для додатків (VBA) можна завантажити надбудову "Пакет аналізу VBA". Для цього необхідно виконати ті ж дії, що і для завантаження надбудови "Пакет аналізу". У вікні **Доступні надбудови** встановіть прапорець поряд з елементом **Пакет аналізу VBA.**

Список доступних інструментів наведений нижче.

- F -тест для дисперсії по двох вибірках
- T -тест
- Z -тест
- Аналіз Фур'є
- Вибірка
- Генерація випадкових чисел
- Гістограма
- Дисперсійний аналіз
- Коваріація
- Кореляція
- Описова статистика
- Ранг і персентиль
- Регресія

- Ковзаюче середнє
- Експоненціальне згладжування

Примітка Функції аналізу даних можна застосовувати тільки на одному листі. Якщо аналіз даних проводиться в групі, що складається з декількох листів, то результати будуть виведені на першому листі, на інших листах будуть виведені порожні діапазони, формати, що містять тільки. Щоб провести аналіз даних на усіх листах, повторите процедуру для кожного листа окремо.

4.2 Використання спеціалізованих статистичних пакетів програм

Для глибокого та професійного використання емпіричних методів програмної інженерії доцільно використовувати професійні статистичні пакети програм такі, як STATISTICA, SPSS та ін. Обмежимось коротким оглядом цих пакетів, оскільки вони комерційні, що затрудняє їх використання у навчальному процесі.

STATISTICA — пакет для всестороннього статистичного аналізу, розроблений компанією StatSoft [17]. В пакеті *STATISTICA* реалізовані процедури для аналізу даних (data analysis), управління даними (data management), добування даних (data mining), візуалізації даних (data visualization).

Система STATISTICA складається з окремих модулів, кожний з яких є повноцінним Windows-застосунком. Можна швидко і зручно переключатися з одного модуля в інший, клацаючи мишею на значках модулів на робочому столі чи активізуючи відповідне вікно застосунку (якщо воно вже було відкрите) або вибираючи модулі в діалоговому вікні. Інтерфейс системи може бути вбудований у конкретний проект користувача.

Найсильнішою стороною пакета є графіка і засоби редагування графічних матеріалів. Представлено сотні типів графіків, матриці і піктограми. Існує можливість розробити свій дизайн графіка і додати його до меню. Засоби керування графіками містять у собі роботу одночасно з декількома графіками, зміну розмірів складних об’єктів, розширені можливості малювання з додаванням художньої перспективи і спеціальних ефектів, розбивку сторінок.

SPSS Statistics (абревіатура англ. "Statistical Package for the Social Sciences" - "статистичний пакет для соціальних наук") - комп'ютерна програма для статистичної обробки даних, один з лідерів ринку в області комерційних статистичних продуктів, призначених для проведення прикладних досліджень в соціальних науках [18].

Між 2009 і 2010 назва програмного забезпечення SPSS була змінена на PASW (Predictive Analytics SoftWare) Statistics.

28 липня 2009 компанія оголосила, що вона була придбана компанією IBM за 1,2 млрд дол. США. За станом на січень 2010 року компанія стала називатися "SPSS: An IBM Company".

На думку деяких авторів, SPSS "займає провідне положення серед програм, призначених для статистичної обробки інформації".

Норман Най, Хедли Халл і Дейл Бент розробили першу версію системи в 1968 році, потім цей пакет розвивався у рамках університету Чикаго. Перше

призначене для користувача керівництво вийшло в 1970 році у видавництві McGraw - Hill, а з 1975 року проект виділився в окрему компанію SPSS Inc. Перша версія пакету під Microsoft Windows вийшла в 1992 році. На даний момент також існують версії під Mac OS X і Linux.

У 2009 році компанія SPSS виробила ребрендинг свого статистичного пакету, який тепер став називатися PASW Statistics (Predictive Analytics SoftWare). 29 липня 2009 року компанія SPSS оголосила[5] про те, що вона отримується фірмою IBM.

Можливості:

- Введення і зберігання даних.
- Можливість використання змінних різних типів.
- Частотність ознак, таблиці, графіки, таблиці зв'язаності, діаграми.
- Первинна описова статистика.
- Маркетингові дослідження.
- Аналіз маркетингових досліджень.

4.3 Загальні відомості про мову R

Зупинимось детальніше на спеціалізованій мові програмування, що більше підходить для використання у емпіричних дослідженнях з програмної інженерії, які проводяться професійними програмістами [18]. R – це об'єктно орієнтована мова та середовище програмування для статистичного аналізу даних. Вона була створена на початку 1990 р. науковцями Оклендського університету (Нова Зеландія) Россом Іхакою (Ross Ihaka) та Робертом Джентльменом (Robert Gentleman) на основі мови S. На відміну від багатьох інших засобів статистичного аналізу R є **вільно розповсюджуваним програмним забезпеченням** з відкритим вихідним кодом. Її особливістю також є те, що вона може використовуватися у всіх операційних системах, зокрема у Windows, Linux, Mac OS тощо.

В 1997 р. був створений CRAN (Comprehensive R Archive Network (http://cran.r-project.org), який є репозиторієм, що містить систему R, бібліотеки, матеріали та інші пов'язані з R ресурси (рис. 4.1).

З цього сайту можна завантажити потрібну версію R, а також пакети з додатковими бібліотеками.

Сьогодні R використовують такі лідери світової економіки, як Google, Pfizer, Merck, Bank of America, InterContinental Hotels Group, Shell та інші. Засоби інтегрування з R мають найбільш популярні комерційні пакети статистичного аналізу даних Statistica, SPSS та SAS.

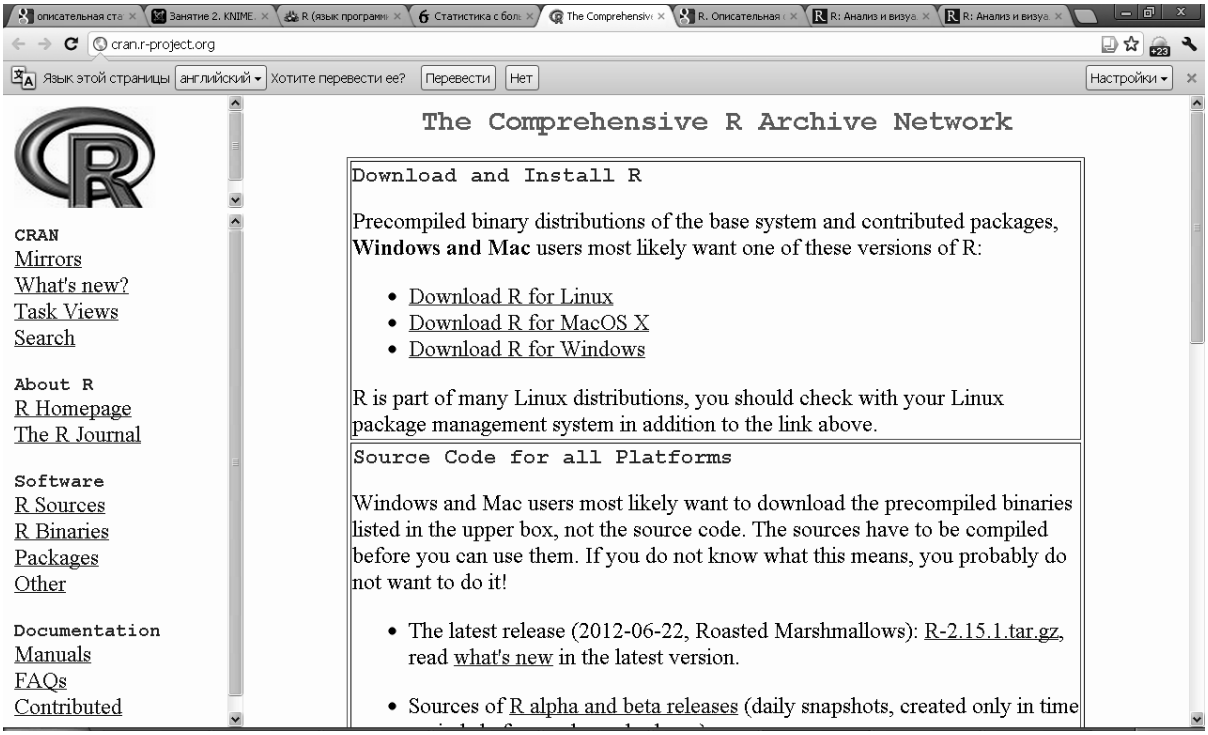

Рис. 4.1. Сайт CRAN

При запуску R виводиться повідомлення (рис. 4.2).

В наступному рядку стоїть символ "> ", який є запрошенням для введення команд.

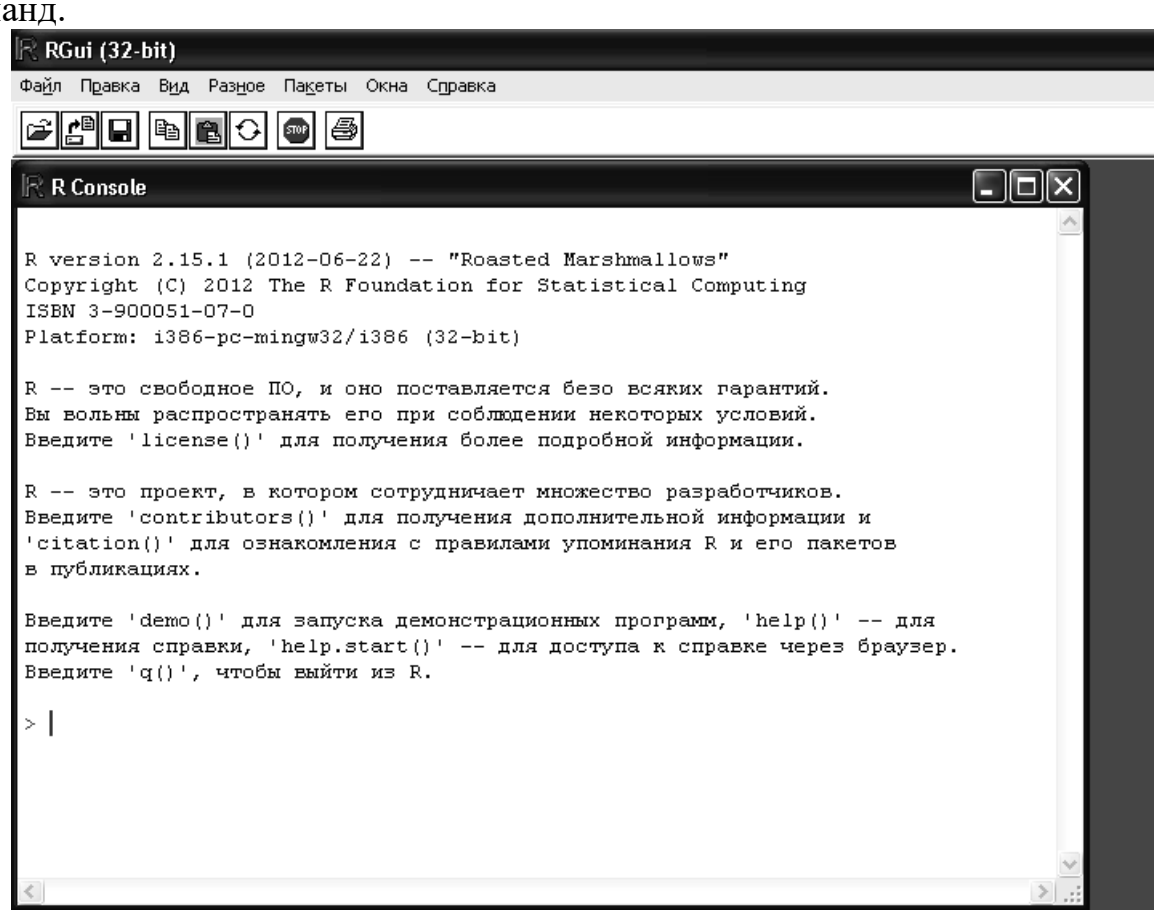

Рис. 4.2. Початкове повідомлення R

Особливістю мови R є те, що різниця між великими та малими літерами є суттєвою, тобто вони сприймаються як різні символи.

В мові R найпростішою структурою є вектор, який містить одне або декілька значень досліджуваної вибірки. На основі векторі можна створювати складніші структури – матриці, масиви тощо. Скалярні величини в R подають як вектори, що містять одну компоненту.

Елементи вектора можуть належати до одного з таких типів даних: numeric, integer, character, complex, logical.

Numeric (еквівалентні позначення double, real). Об'єкти цього типу можуть містити тільки цілі або дробові числа і використовуються для виконання математичних операцій. За допомогою функції is.numeric () можна перевірити, чи належить об'єкт до цього типу.

За замовчанням програми та дані зберігаються та шукаються у робочій директорії. Визначити робочу директорію можна за допомогою команди:

> getwd().

Змінити робочу директорію можна за допомогою команди:

Робота з командним рядком. Основне робоче середовище системи R - командний рядок. Тому при запуску системи ви побачите коротку інформацію про роботу з командним рядком і запрошення введення. Ви завжди можете вийти, ввівши команду q(). Ця команда є викликом функції з ім'ям "q" і без аргументів. Ця функція виробляє вихід з системи.

Для того, щоб отримати довідку про деяку функцію, треба викликати функцію help(<имя-функции>) з цим ім'ям як аргумент, в лапках або без (наприклад, help("help") або help(help)). Аналогом цієї функції є команда ?<имя-функции>, але в цьому випадку ім'я функції приводиться без лапок (наприклад, ?help). Обидва ці способи приведуть до одного і того ж - у разі наявності довідки по цій функції буде відкрито нове вікно (створена нова термінальна сесія, якщо використовується версія для UNIX) з вмістом цієї довідки. Вийти з перегляду довідки можна, натиснувши клавішу "q".

Послідовність команд R можна вводити у файл з наступним завантаженням цього файлу як цілісного сценарію. Таке завантаження здійснюється функцією source(<имя-файла>).

Основні типи даних в R. Основним типом даних в системі R є число. Числа представляються звичайним для мов програмування чином - арабські цифри, десятковий дріб розділяється точкою, маленькі і великі числа можна означати у виді - 0.4+E3. При введенні як команда числа, в результаті буде виведено воно ж.

Вектор - це набір декількох чисел. Кількість чисел називається завдовжки вектору. Формується вектор функцією c(<число1>, <число2>, ..), якщо ввести її послові запрошення R, то виведеться значення вектору :

> c(1, 2, 3, 4, 5)

[1] 1 2 3 4 5

Такий же вектор, що складається з послідовних чисел, можна отримати і командою <начальное-число>:<конечное-число> (наприклад, 1:5). Така команда дозволяє отримувати вектор з послідовних цілих чисел від початкового до кінцевого включно.

Окремо від усіх чисел стоять псевдочисла NaN (Not a number) і NA (Not assigned). Перше є результатом выражений-неопределенностей (0/0, ∞ − ∞), а друге використовується як синонім невизначеного доки значення (аналог NULL в системах управління базами даних).

Команда:

> x = seq(a, b, len=n)

створює вектор, що містить n чисел, які збільшуються від a до b з постійним кроком(n - 1). Наприклад:

```
> x = seq(- 3, 3, len=7)
> x
[1] - 3 -2 -1  0  1  2  3
```

Команда:

> > y <- seq(from = a, to = b, by = c)

створює вектор, що містить послідовність чисел, які збільшуються від a до b з постійним кроком c. Наприклад:

```
> y <- seq(from = 2, to = 20, by = 2)
> y
 [1]  2  4  6  8 10 12 14 16 18 20
```

Вектор довжиною n*m мі можемо перетворити у матрицю за допомогою функції:

mat=matrix(data = y, nrow = n, ncol = m).

Наприклад:

```
> y <- seq(from = 2, to = 40, by = 2)
> mat=matrix(data = y, nrow = 4, ncol = 5)
> mat
     [1] [,2] [,3] [,4] [,5]
[1,]    2   10   18   26   34
[2,]    4   12   20   28   36
[3,]    6   14   22   30   38
[4,]    8   16   24   32   40
```

Ми також можемо об' єднати два вектори до матриці, наприклад:

```
> x <- seq(from = 2, to = 20, by = 2)
> y <- seq(from = 22, to = 40, by = 2)
> mat=matrix(data = c(x, y), nrow = 10, ncol = 2)
> mat
```

```
      [1] [,2]
 [1,]   2   22
 [2,]   4   24
 [3,]   6   26
 [4,]   8   28
 [5,]  10   30
 [6,]  12   32
 [7,]  14   34
 [8,]  16   36
 [9,]  18   38
[10,]  20   40
```

Ще один шлях створення матриці полягає в тому, щоб задати матрицю, наприклад:

Змінні в R. У мові системи R є можливість зберігати дані в змінних. Змінні можуть мати імена, що складаються з символів латинського алфавіту, арабських цифр, символів підкреслення і точки. Імена можуть починатися тільки з алфавітного символу або символу точки, але в останньому випадку після точки може йти тільки алфавітний символ. Імена змінних чутливі до регістра.

Привласнення змінних здійснюється двома способами - звичним багатьом програмістам знаком = і не таким звичним в цій якості знаком "стрільця" <-:

```
> x = 7
> x
[1] 7
> x <- 8
> x
[1] 8
```

В даному випадку команда, що складається просто з імені змінної, виводить значення цієї змінної. У системі R так само легко вивести значення змінних усіх типів, і це дозволяє в будь-який час відстежувати такі значення, не удаючись до серйозних методів відладки.

Векторам можна привласнювати значення, відповідні не лише числам, але і, наприклад, векторам (x = c(1, 2, 3)), а також результати виразів (x = c(1, 2, 3) + c(2, 3, 4)).

Доступ до елементів вектору можна отримати, вказавши номер цього елементу в квадратних дужках після вектору (змінною, що містить вектор, або вираження, результатом якого буде вектор) :

```
> x = 1:20 # 1-й випадок
> x[5]
[1] 5
> (1:20)[5] # 2-й випадок
[1] 5
> 1:20[5] # 3-й випадок
Помилка в 1:20[5] : NA/NaN -аргумент
```

У третьому випадку система вважала, що індекс 5 застосовується до 20, що і викликало помилку. Правильний приклад вказаний в другому випадку, вираження 1:20 потрібно узяти в дужки.

Також можна гнучкіше адресувати елементи вектору при використанні квадратних дужок. Можна вибрати відразу декілька елементів, що цікавлять, вказавши в квадратних дужках вектор, що містить індекси необхідних елементів :

```
> x = 2 * c(1, 2, 3, 4, 5)
> x[c(1, 3, 5)]
[1] 2 6 10
```

Замість вказівки конкретних індексів, в квадратних дужках можна вказати умову.

Відповідно, будуть відібрані тільки ті елементи, які задовольняють цій умові, :

```
> x[x > 5]
[1] 6 8 10
```

Як і у випадку з поодиноким індексом, вираженню з множинним індексом можна привласнювати значення - у такому разі значення привласнюються тільки вказаним елементам:

```
> x[c(2, 4)] = 0
> x
[1] 2 0 6 0 10
> x[x > 5] = 0
> x
[1] 2 4 0 0 0
```

При привласненні змінної значення, якщо цієї змінної не було, вона створюється. В результаті може бути створені багато змінних, не усі з яких важливі. Проглянути список створених змінних можна, викликавши функцію ls(). Якщо деякі із змінних вже не потрібні, їх можна видалити функцією rm(<змінні>). Це зручно, якщо значення змінних треба зберігати між сесіями роботи, і зайві змінні тільки заважатимуть.

Операції над основними типами даних

Над числами і векторами можна виробляти звичайні арифметичні операції, і позначаються вони досить традиційно:

- складання: 2 + 3;
- віднімання: 1 - 2;
- множення: 5 * 2;
- ділення: 5 / 2;
- піднесення до степеня: 2 ^ 8.

У разі проведення операції над векторами однакової розмірності, ця операція виробляється почленно над кожною парою елементів векторів-учасників, і результатом є вектор, що складається з результатів цієї дії. У разі, коли розмірність векторів не співпадає, виробляється приведення довжини "коротшого" вектору до довжини "довгого" вектору. Приведення виконується таким чином: вектор придбаває довжину, таку ж, як і у "довшого", і вміст цих

нових елементів береться з вмісту старих елементів вектору, повторюваних циклічно, :

> c(1, 2, 3) + c(2, 2, 2, 2, 2, 2, 2)
[1] 3 4 5 3 4 5 3

При цьому видається попередження:
In c(1, 2, 3) + c(2, 2, 2, 2, 2, 2, 2) :
довжина більшого об'єкту не є твором довжини меншого об'єкту

> 1 + c(1, 3)
[1] 2 4

У другому випадку число 1 представляється як вектор одиничної довжини, і він приводиться до довшого вектору - учасника вираження.

У виразах можуть брати участь звичні функції sin(x), cos(x), tan(x), sqrt(x) і деякі інші. Їх аргументами можуть бути як числа, так і вектори.

Окреме місце в мові R займають вбудовані функції роботи з векторами. До них відносяться:

sum(v) : знаходження суми елементів вектору v;
max(v) : знаходження максимального елементу вектору v;
min(v) : знаходження мінімального елементу вектору v;
prod(v) : знаходження твору елементів вектору v;
length(v) : знаходження кількості елементів вектору v;
mean(v) : знаходження середнього (оцінки математичного очікування) вектору v;
var(v) : знаходження варіації (оцінки дисперсії) вектору v;
sort(v) : повернення вектору з такою ж довжиною, як і у v, елементи якого відсортовані в порядку зростання.

За допомогою цих функцій операції з векторами набирають вигляду простих арифметичних виразів, на відміну від громіздких циклів, що виконують ті ж функції, в мовах програмування загального призначення. Так, функцію mean(v) можна замінити вираженням sum(v)/length(v). В результаті отримувані вирази стають нескладними у відладці.

> setwd(“…”)

У лапках необхідно вказати шлях до потрібної директорії. В RGui це можна зробити також, обираючи в Меню: Файл – Змінити папку.

Для створення програм доцільно використовувати скрипти. Їх можна створювати за допомогою будь-якого текстового редактора, а також у спеціальному вікні, що з’являється, якщо обрати в головному меню: Файл – Новий скрипт. Для виклику вже готових скриптів необхідно обрати у головному меню: Файл – Відкрити скрипт. При створенні скриптів слід звернути увагу на те, що в них не використовують позначки запрошення “>” та “+”, що з’являються на початку нового рядка при роботі у консолі.