ОСНОВИ РОБОТИ З R.
ОБРОБКА СТАТИСТИЧНИХ ДАНИХ.

Мета роботи. Ознайомитися з інтерфейсом RStudio, навчитися працювати в режимі консолі і шляхом написання скриптів, а також підключати зовнішні пакети, вивчити основні методи обробки статистичних даних.

Завдання.

1. Відкрити середовище RStudio для програмування на мові R та обробки статистичних даних, використовуючи менюПуск.

2. Виконати усі командні рядки з методичних вказівок та отримати такі жрезультати.
   При цьому вводити ті рядки, які йдуть після знаку >, без цього знаку рядок буде з'являтися як результат.

3. Зберегти виконану роботу у файлі Ry папці, яку треба створити на дискуD:\ з ім'ям своєї групи, у випадку дистанційного навчання відправити його викладачу на поштуLarisaK2010@ukr.net.

Вказівки до виконання роботи.

1. Консольний режим.

Після запуску RStudio користувач потрапляє в консольний режим роботи (рис. 1). Будь-яка команда, написана користувачем, буде відразу виконана R після натискання Enter.

Розглянемо основні вирази в R: числа, рядки і логічні змінні. Можна використовувати R як калькулятор, наприклад:

```
> 1 + 1
[1] 2
> 6 * 7
[1] 42
> Sqrt (16)
[1] 4
```

Результат відразу з'явиться в консолі. Рядки друкуються в лапках: подвійних або одинарних:

```
>"Hello world!"
[1] "Helloworld!"

> 'Hello world!'
[1] 'Hello world!'
```

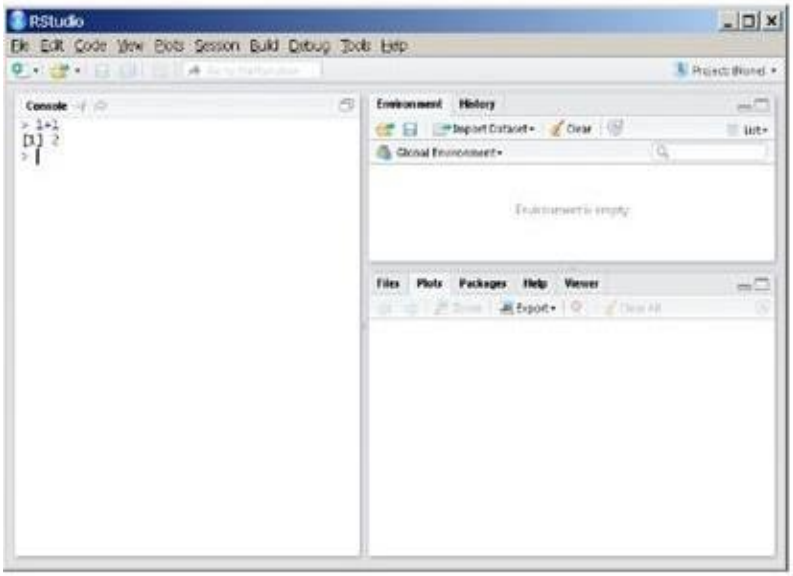

Рис. 1 . Консольний режим в RStudio.

Логічні вирази повертають TRUE або FALSE

```
> 3 < 4
      [1] TRUE
```

Щоб порівняти два вирази, використовується подвійний знак рівності:

```
> 2 + 2 == 5
      [1] FALSE
```

Як і в інших мовах програмування, можна зберігати значення в змінну.

Збережемо 42 в змінну x:

```
>x<-42
```

І в зворотну сторону:

```
>5 ->X
```

Можна роздрукувати значення змінної в будь-який час, просто набравши її ім'я в консолі.Спробуємо надрукувати поточне x:

```
>x
[1]  42
```

Можна так само повторно призначити будь-яке значення змінної у будь-який час. R чутливий до регістру: змінні – x і X це різні змінні

```
>X
[1] 5
```

Щоб викликати функцію, потрібно звернутися до неї по імені, вказавши в дужках потрібні аргументи. Наприклад, функція суми:

```
>sum (1, 3, 5)
[1]  9
```

Отримати допомогу по функції можна командою help (functionname) або?functionname. У правому нижньому кутку на вкладці Help з'явиться довідка(рис.2):

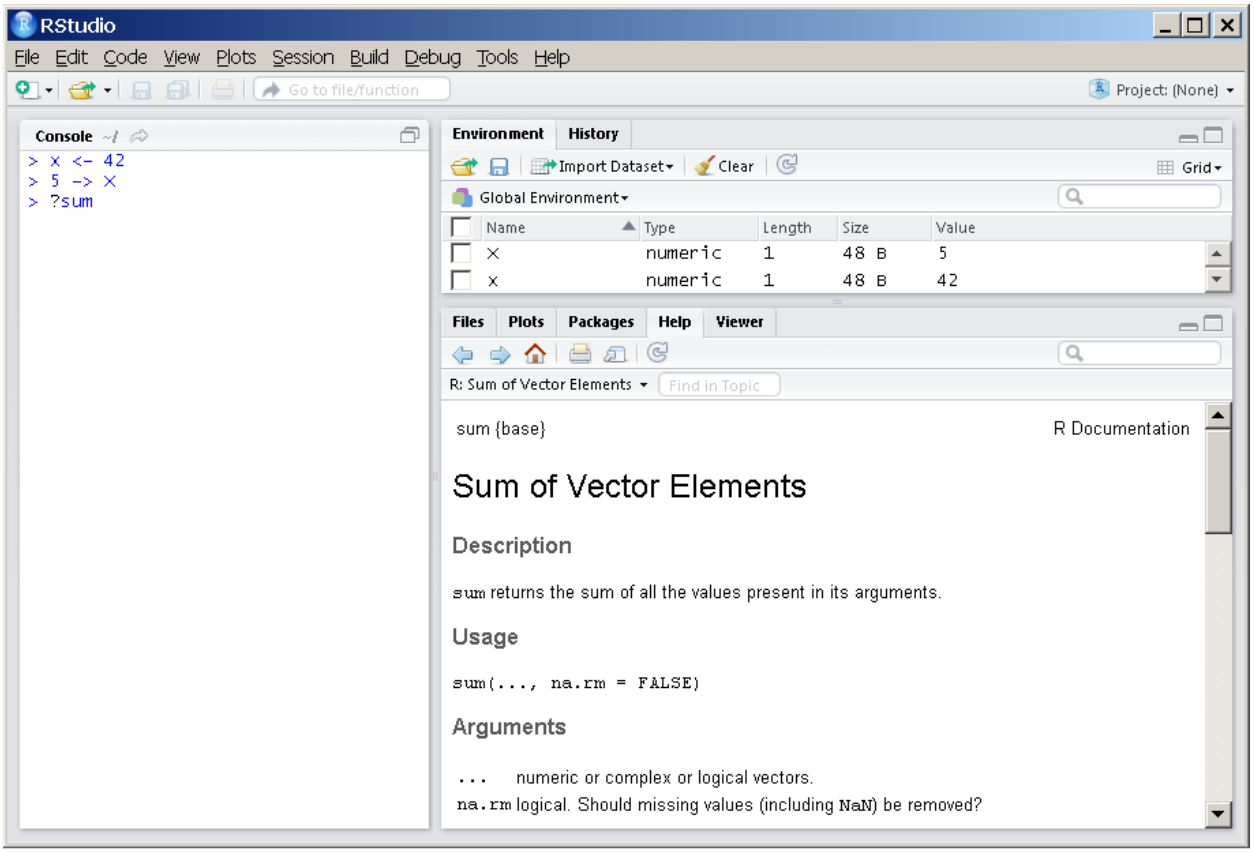

Рис. 2. Довідка з функції sum.

Задамо вектор за допомогою функції с (скор. Від англ. Combine):

```
> y <- c (-3, 2, NA,5)
> y
[1] -3 2 NA 5
```

NA- це пропущене спостереження (від англ. NotAvailable).

Його не слід плутати з NaN (Not a Number - "не число», невизначеність):

```
> 0/0
[1] NaN
```

Спробуємо підсумувати елементи вектора y:

```
> sum (y)
[1] NA
```

Необов'язковим аргументом функції sum є na.rm (скор. від англ. Remove NA), за замовчуванням рівний FALSE.
Якщо вказати для нього значення «істина», то функція суми буде складати всі елементи вектора, виключаючи пропущені:

```
> sum (y, na.rm= TRUE)
[ 1] 4
```

Послідовність чисел можна задати двома способами: start : end або функцією seq ():

```
> 5: 9
[1] 5 6 7 8 9
> seq (5,9)
[1] 5 6 7 8 9
> seq (10,50, by = 10)
[1] 10 20 30 40 50
```

Звертатися до елементів вектора можна, використовуючи квадратні дужки:

```
>sentence <- c ( 'mack', 'the', 'knife')
> sentence [3]
[1] "knife"
> sentence [c (1,3)]
[1]“mack"
```

Або можна задати елементам вектора імена:

```
> ranks <- 1:3
> names (ranks) <- c ("first", "second", "third")
> ranks
```

```
firstsecond  third
   1        2     3
>ranks["first"]
first

1
```

Написання скриптів.

В R зручніше писати не по одній команді, а відразу цілий набір команд і потім запускати їх все на виконання. Для цього потрібні скрипти.
Щоб створити скрипт, слід вибрати File -> New File -> R Script.

Відкриється нова область, в якій можна писати команди. Коментарі, які не виконуватиме R, пишуться зі знака # (рис. 3).

При натисканні Enter команди у скріпті виконуватися не будуть, а буде лише здійснений перехід на новий рядок.

Щоб виконати команду в режимі скрипта, слід поставити курсор на потрібний рядок і натиснути Ctrl + Enter.

Якщо команда займає більше одного рядка, то необхідно ставити знак + в кінці кожного рядка. Команда виконується по рядках, у кожному рядку потрібно натискати Ctrl + Enter, або виділити всю команду і натиснути Ctrl + Enter один раз.

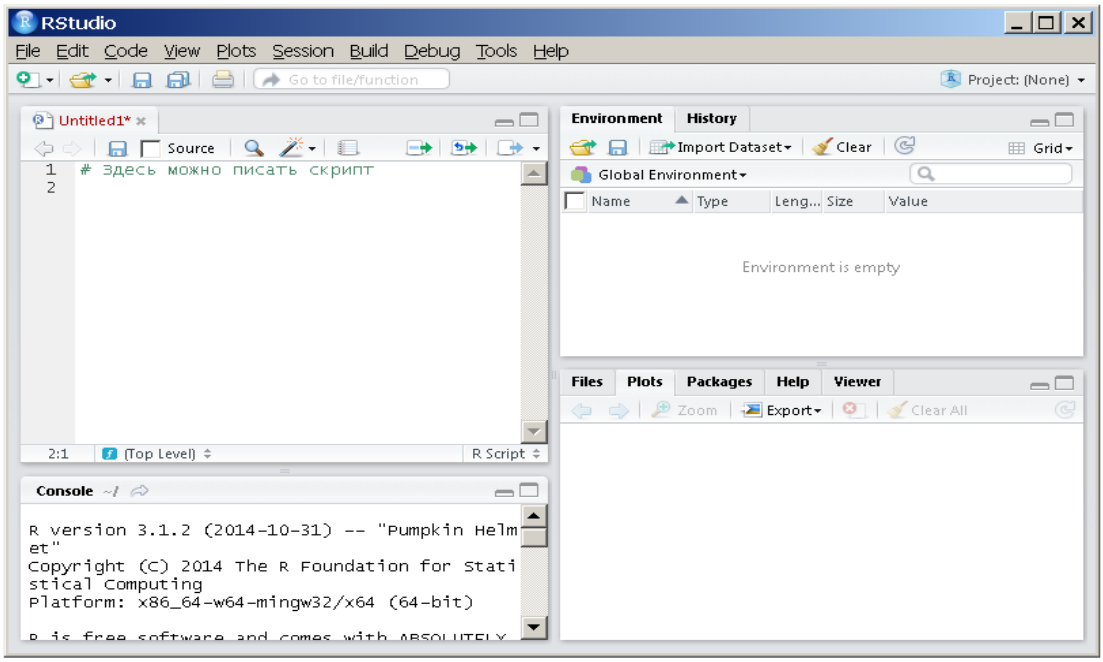

Рис. 3. Створення нового скрипта.

R може підказувати, які команди доступні, якщо почати вводити перші символи і натиснути або Tab, або Ctrl + Enter.

В основному в R працюють з наборами даних. Така структура має вRназвydata.frame та являє собою таблицю, де кожний стовпчик – це деяка змінна а кожний рядок – це деяке спостереження.

Створимо у режимі скрипта data.frame. Нехай є спостереження за зростом і вагою деяких людей. Задамо два вектора:

```
>rost <- c (160, 175, 155, 190, NA)
>ves <- c (NA, 70, 48, 85, 60)
```

І об'єднаємо їх в набір даних, який помістимо в пере- менную df, а потім виведемо на екран:

```
>df<- data.frame (rost,ves)
>df
```

В консолі отримаємо наступну таблицю:

```
    rost ves
    1 160 NA
    2 17570
    3 15548
    4 19085
    5 NA 60
```

Звертатися до конкретних спостереженнями df можна, використовуючи квадратні дужки :

```
    > df [3,1]
    [1] 155
```

Звертатися до змінних можна, використовуючи знак $ або вказуючи на стовпець з пропуском номера рядка:

```
    > df$rost
    [1] 160 175 155 190 NA
```

або

```
> df [ ,1]
[1] 160 175 155 190 NA
```

Звертатися до спостережень можна , вказуючи конкретний рядокі пропускаючи номер стовпчика

```
> df [4, ]
rostves
4 190 85
```

Основні описові статистики (середнє, стандартневідхилення і медіану) можна отримати за допомогою функцій mean, sd і median:

```
mean (df$rost, na.rm = T)
[1] 170
sd (df$rost, na.rm = T)
[1] 15.81139
median (df $ rost, na.rm = T)
[1] 167.5
```

Можна зберегти скрипт, натиснувши File -> Save. При першому зберіганні R запропонує вибрати кодування. Рекомендується вказати UTF-8 (рис.4), щоб російські букви (наприклад, у коментарях) відображалися коректно. Потім необхідно вибрати директорію і задати ім'я файлу, який буде збережений з розширенням * .R.

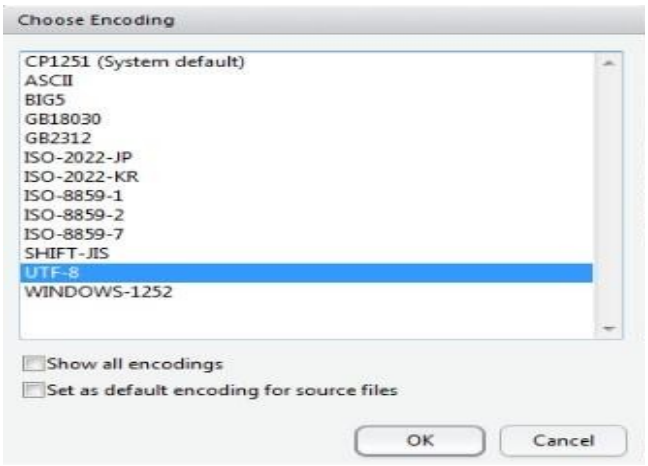

Рис. 4. Вибір кодування при збереженні.

Установка пакетів.

В R існує базовий набір пакетів (бібліотек), який містить самі необхідні функції [2]. Для реалізації завдань економетрікі потрібні додаткові пакети.
Для роботи з зовнішніми пакетами необхідно виконати дві дії - встановлення та підключення потрібного пакету. Установку неободно виконати один раз, а підключати у кожній робочій сесії.
Для установки пакетів існує функція install.packages. Також автоматично будуть доустановлени пов'язані пакети.
Для підключення встановленого пакета слід скористатися функцією library [1]. Наприклад, встановимо такі пакети:
psych - містить функції для розрахунку описових статистик.
dplyr - містить функції для роботи з data.frame;
ggplot2 - найпотужніший пакет для побудови красивих графіків, діаграм, карт і т. д.

```
install.packages (c ( "psych", "dplyr", "ggplot2"))
```

Прямим повідомленням про помилку установки є тільки слово Error, що з'являється в консолі. Всі інші повідомлення Warning є просто попередженнями про що-небудь.
Для того щоб визначити, які пакунки є необхідними для роботи, можна скористатися пошуком в мережі Інтернет, задавши питання на англійській мові. Наприклад, щоб знайти, в якому пакеті знаходиться алгоритм Левенберга-Марквардта для розрахунку нелінійного МНК можна набрать в поиске «levenberg-marquuardt algorithm in r» та першим посиланням буде пакет minpack.lm вR.
Для виконання даної роботи знадобиться підключити наступні пакети:

```
library ( "psych") # описові статистики
library ( "lmtest") # тестування гіпотез в лінійних моделях
library ( "ggplot2") # графіки
library ( "dplyr") # маніпуляції з даними
library ( "MASS") # підгонкарозподілів
```

Отримаємо, наприклад, опис набору даних по автомобілям cars командою:

```
help (cars)
```

Результат виконання команди (в правому нижньому кутку на вкладці Help) показаний на рисунку 5.
У цьому наборі даних 50 спостережень і дві змінні (швидкість, миль / годину і довжина гальмівного шляху в футах).

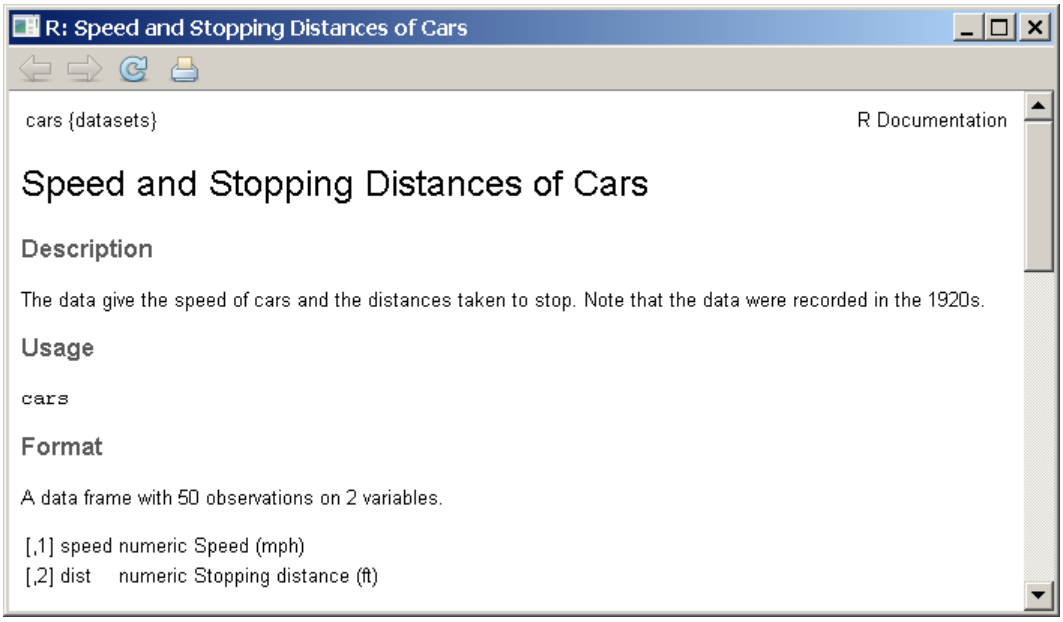

Рис. 5. Довідка з набору даних cars.

Помістимо у змінну d вбудований в R набір даних по автомобілям:

```
d <- cars # цей набір даних знаходиться в базовому пакеті datasets
```

Тепер d має тип data.frame (набір даних), у чьому можна впевнитися, подивившись у правому верхньому куті вікна таблиці середовища Environment(рис. 6).
Для цього повинен бути обраний режим Grid. За допомогою кнопкиможна переглянути вміст наборуданих у вигляді табліці.

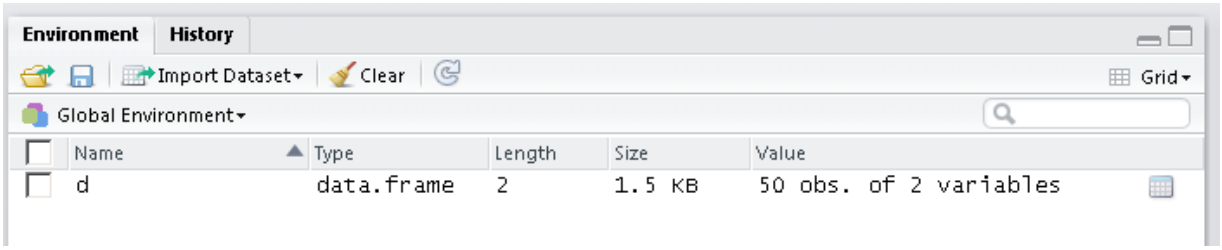

Рис. 6. Опис набору даних d в таблиці середовища Environment.

Наступною командою можна подивитися на цей набір даних, в результаті чого будуть перераховані всі змінні і типи даних:

```
>glimpse (d) # функція з пакета dplyr
```

Результат виконання команди з'явиться в консолі:

```
> d <- cars
> glimpse(d)
```

Observations: 50
Variables: 2

$ speed (dbl) 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, …
$ dist (dbl) 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, ...

Змінні speed і dist мають тип даних dbl (double) та містить по 50 спостережень. Для інших типів даних використовуються такі скорочення:chr (character / string), int (integer), fctr (factor), tims (time), lgl (logical).

Подивимося на перші шість спостережень набору даних d :

>head (d) # функція з базового пакету utils
>speed dist
1 4 2

2 4 10

3 7 4

. 4 722

5 816

6 910

і останні шість спостережень

> tail(d) #функція з базового пакету utils
speed dist

45 2354

46 2470

47 2492

48 2493

49 24 120

50 25 85

Отримаємотаблицю з описовими статистиками: середнє, мода,медіана, стандартне відхилення, мінімум / максимум, асіметрія, ексцес і т. д.:

>describe(d) # функция из пакета psych

Vars	n	mean	sd	median	trimmed	mad	min	max	range	
speed	1	50	15.40	5.29	15	15.47	5.93	4	25	21
dist	2	50	42.98	25.77	36	40.88	23.72	2	120	118

Побудуємо гістограму абсолютних частот для змінної dist (довжини гальмівного шляху). Скористаємося функцією qplot, задавши джерело даних d (аргумента data), змінну для побудови графіка (dist), підпишемо осі (параметри функцій xlab, ylab) та назву графіка (параметр main).

```
# функция из пакета ggplot2

>qplot(data=d,dist,xlab="Длинатормозногопути(футы)",ylab="Число …
автомобилей", main="Данныепоавтомобилям1920-х")
```

Результат виконання даної функції надано на рисунку 7.

Рис. 7. Гістограма абсолютних частот для змінної dist.

Можна побудувати також діаграму щільності розподілу (рисунок 8):

```
# функция из базового пакета graphics
>hist(d$dist,probability=TRUE, col="grey")
```

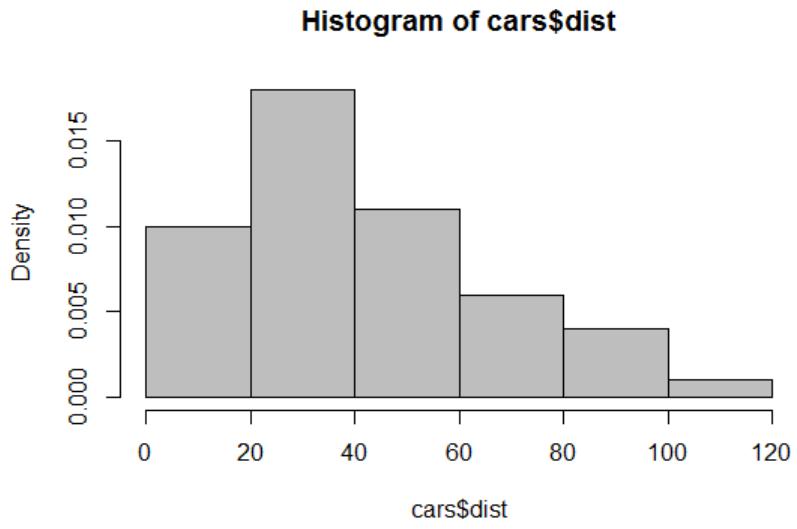

Рис. 8. Гістограма щільності розподілу для змінної dist.

Підгонимо щільність розподілу Вейбула, розташувавши результат (оцінки параметрів розподілу) у змінній fit:

```
fit<-fitdistr(d$dist,"weibull") # функція з пакету MASS
```

Змінна fit тепер буде представлена у вигляді списку (List) із п'ятиелементів, цевідображаєтьсяуEnvironment (рис. 9).

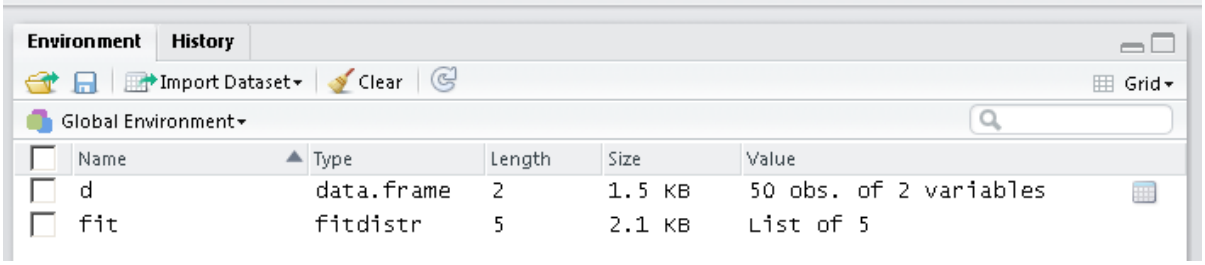

Рис.9. Опис змінної fity таблиці Environmentсередовища R.

Доступ до елементів списку можна отримати за допомогою символу $.
Оцінки двох параметрів розподілу Вейбула були розраховані методом максимальної правдоподібності. Переглянемо їх, звернувшись до елементу списку fit:

```
>fit$estimate
shape            scale
1.7237148.15234
```

Покажимо на тому ж графіку теоретичну щільність розподілу Вейбула (рис. 10).

```
>xvals<-seq(0,120,.20) #значенияпоосиабсциссот0до120с шагом0,2

>lines(xvals,dweibull(xvals,shape=fit$estimate[1], scale=fit$estimate[2]))
```

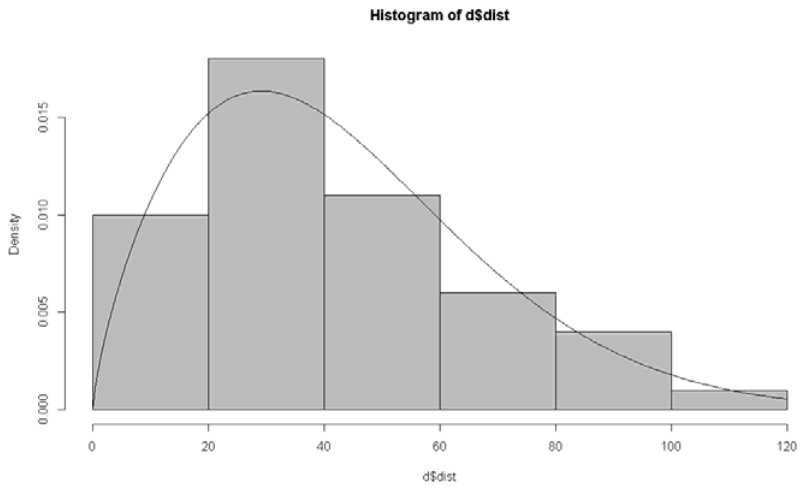

Рис. 10. Гістограма розподілу змінної dist.

Перший аргумент функції **lines**– це значення по осі абсцис, на основі яких буде побудовано графік.
Далі вказується функція щільності **dweibull**. Для неї потрібно вказати значення аргументу для розрахунку та значення двох параметрів розподілу: коефіцієнт форми **(shape)** та масштабу **(scale)**.

Когтрольні запитання.

1. Як знайти середнє вибіркове значення та стандартне відхилення ?
2. Як перевірити на явність пропусків у вихідних даних ?

Варіанти завдань.

№ варианта	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**
Набор данных	CO2		ChickWeight		Orange		airquality		faithful	
Имя переменной (вектора)	conc	uptake	weight	Time	age	circumference	Wind	Temp	eruptions	waiting