Course: R Language in Computational Probability and Statistics

Lecture 4: Frames. Lists

Lecturer: Nataliia Kruhlova

Списки

Найбільш важливими об’єктами в аналізі даних є списки і фрейми.

Список даних **list** —це упорядкований набір змінних, що об'єднані у спільний об'єкт. Його члени називають компонентами (полями). Компоненти можуть мати різні властивості. Список будується за допомогою функції:
list(name1=object1, name2=object2,...)
Приклад.

```
> Department<-list(Assistant=c("Petrenko","Nikolaenko", "Petrov"),Senior_lecturer=c("Ivanov","Lemeshko"),Associate_professor=c("Kuzmenko","Rudenko"),Age=c(29,31,59,48,54,90,96))
> Department
$Assistant
[1] "Petrenko"   "Nikolaenko" "Petrov"
$Senior_lecturer
[1] "Ivanov"   "Lemeshko"
$Associate_professor
[1] "Kuzmenko" "Rudenko"
$Age
[1] 29 31 59 48 54 90 96
```

До поля списка звертаються, вказавши його номер у подвійних квадратних дужках після імені списку, або у форматі **List_name$Some_object.**
Приклад.

```
> Department[[3]]
[1] "Kuzmenko" "Rudenko"
> Department$Associate_professor
[1] "Kuzmenko" "Rudenko"
```

Об’єкт Department[[3]] є вектором, тому до його першої компоненти можна звернутись декількома способами.

Приклад.

```
> Department[[3]][1]
[1] "Kuzmenko”
> Department$Associate_professor[1]
[1] "Kuzmenko"
```

Більшість функції в R надають свої результати у вигляді списків.

Приклад. Знайдемо власні функції і власні вектори деякої квадратної матриці.

```
> a<-matrix(1:9,nrow=3)
> result<-eigen(a)
> result
eigen() decomposition
$values
[1]  1.611684e+01 -1.116844e+00 -5.700691e-16

$vectors
           [,1]       [,2]       [,3]
[1,] -0.4645473 -0.8829060  0.4082483
[2,] -0.5707955 -0.2395204 -0.8164966
[3,] -0.6770438  0.4038651  0.4082483
```

Зверніть увагу на **mode (result)**, **class (result), str(result).**

```
> mode(result)
[1] "list"
> class(result)
[1] "eigen"
> str(result)
List of 2
 $ values : num [1:3] 1.61e+01 -1.12 -5.70e-16
 $ vectors: num [1:3, 1:3] -0.465 -0.571 -0.677 -0.883 -0.24 ...
 - attr(*, "class")= chr "eigen"
```

Фрейми (набори даних)

Зазвичай набори даних для статистичного аналізу даних мають клас **data.frame**, тобто є системами даних. Змінні, які формують різні стовпчики

можуть відрізнятися за типом, але всі елементи стовпчика повинні бути одного типу і довжини стовпчиків мають співпадати.

Змінні об'єднуються в систему даних (фрейм) таким чином:
d.frame<- data.frame (name1=obj1, name2=obj2,...)

Якщо компоненти списку **(list)** задовольняють умовам системи даних, то список може бути перетворений на систему даних за допомогою функції **as.data.frame().**
До стовпчика фрейму потрібно звертатися так: **d.frame$name1.**

Команда **data(some_data)** додає фрейм в робоче середовище.

Команда **help(some_data)** виводить інформацію про фрейм.

Щоб вивести на екран структуру даних, можна скористатися командою **str().**

Приклад. Дата фрейм **attitude** –вбудований масив даних, який містить рейтинг департаментів одної фінансової компанії. Виведіть дані з 11 по 20 рядок цього фрейму.

```
> attitude[11:20,]
   rating complaints privileges learning raises critical advance
11     64         53         53       58     58       67      34
12     67         60         47       39     59       74      41
13     69         62         57       42     55       63      25
14     68         83         83       45     59       77      35
15     77         77         54       72     79       77      46
16     81         90         50       72     60       54      36
17     74         85         64       69     79       79      63
18     65         60         65       75     55       80      60
19     65         70         46       57     75       85      46
20     50         58         68       54     64       78      52
```

Додати новий стовпчик у набір даних можна за допомогою присвоєння:

Data.frame_name$name_variable<-c(…).

Видалити стовпчик із фрейму можна командою:

Data.frame_name$name_variable<-NULL

Приклад. Потрібно перших 20 рядків фрейму **attitude** зберегти в новий фрейм. В цей набір даних додати нову змінну **suma**, в яку введено суму відповідних елементів стовпчиків **learning, critical, advance**, а дані стовпчику **privileges** видалити.

```
> newdata<-attitude[1:20,]
> newdata$suma<-newdata[,4]+newdata[,6]+newdata[,7]
> newdata$privileges<-NULL
> newdata
   rating complaints learning raises critical advance suma
1      43         51       39     61       92      45  176
2      63         64       54     63       73      47  174
3      71         70       69     76       86      48  203
4      61         63       47     54       84      35  166
5      81         78       66     71       83      47  196
6      43         55       44     54       49      34  127
7      58         67       56     66       68      35  159
8      71         75       55     70       66      41  162
9      72         82       67     71       83      31  181
10     67         61       47     62       80      41  168
11     64         53       58     58       67      34  159
12     67         60       39     59       74      41  154
13     69         62       42     55       63      25  130
14     68         83       45     59       77      35  157
15     77         77       72     79       77      46  195
16     81         90       72     60       54      36  162
17     74         85       69     79       79      63  211
18     65         60       75     55       80      60  215
19     65         70       57     75       85      46  188
20     50         58       54     64       78      52  184
```

За допомогою функції **data.frame()** утворюється набір даних з окремих векторів-змінних, які перетворюються на стовпчики фрейму.

Приклад.

```
> Names<-c("Ivanov","Petrenko","Sidorenko","Zmienov","Poltavsky")
> Math<-c(76,90,63,56,89)
> History<-c(76,88,92,85,59)
> Algebra<-c(67,85,74,69,91)
> Sertificate<-c("Y","N","N","Y","N")
> Candidates<-data.frame(Math,History,Algebra,Sertificate,row.names=Names)
> Candidates
          Math History Algebra Sertificate
Ivanov      76      76      67          Y
Petrenko    90      88      85          N
Sidorenko   63      92      74          N
Zmienov     56      85      69          Y
Poltavsky   89      59      91          N
```

Параметр **row.names** виводить назви об'єктів (рядків набору даних). Іменами змінних стають назви векторів стовпців, з яких створено фрейм. У разі потреби назви рядків і стовпців можна переглянути і змінити за допомогою функцій **rownames(), colnames()** і **names()**. Параметр **stringsAsFactors** вказує, чи потрібно перетворювати вектори символьних рядків у змінні типу фактор. За замовчуванням виконується таке перетворення, тому параметр **stringsAsFactors = FALSE** дозволяє і далі працювати із символьними змінними. Для перегляду та, при необхідності, виправлення фреймів застосовується функція **edit()**. До даних, що містяться у наборі даних, можна звертатись так само, як і до елементів матриць.

Приклад.

```
> Candidates[,2]
[1] 76 88 92 85 59
> Candidates[3,]
          Math History Algebra Sertificate
Sidorenko   63      92      74          N
> Candidates[,"Algebra"]
[1] 67 85 74 69 91
```

До змінних з набору даних можна також звертатись, використовуючи формат **Frame$Object.**

Приклад. Потрібно вивести вектор значень змінної **Algebra** набору даних **Candidates**.

```
> Candidates$Algebra
[1] 67 85 74 69 91
```

Щоб перевірити тип змінної у фреймі, використовують функції **is.numeric**, **is.character, is.logical** і т.п. Результатом таких функцій буде логічна змінна зі значеннями TRUE або FALSE.

Приклад.

```
> is.character(Candidates$History)
[1] FALSE
> is.factor(Candidates$Sertificate)
[1] TRUE
```

Збереження і видалення змінних із систем даних

Це можна зробити так само, як і для матриць.

Приклад.

```
> Candidates_new<-Candidates[,c(1,3)]
> Candidates_new
          Math Algebra
Ivanov      76      67
Petrenko    90      85
Sidorenko   63      74
Zmienov     56      69
Poltavsky   89      91
> Candidates2<-Candidates[,-4]
> Candidates2
          Math History Algebra
Ivanov      76      76      67
Petrenko    90      88      85
Sidorenko   63      92      74
Zmienov     56      85      69
Poltavsky   89      59      9
```

Для вибору частини набору даних можна застосовувати функцію **subset.**

>a<- subset (A, subset, select)

A - об'єкт, з якого вибирається підмножина; аргумент **subset -** логічний вираз, для вибору рядків, що залишаються; **select -** логічний вираз, що вказує стовпчики, які залишаються.

Приклад. У фреймі **Candidates** нам потрібно вибрати тільки ті значення змінних **Math, Algebra, History,** які відповідають значенню змінної **Sertificate "Y"**.

```
> A<-subset(Candidates,select=c(1,3,2),Sertificate=="Y")
> A
        Math Algebra History
Ivanov    76      67      76
Zmienov   56      69      85
```

Зауважимо, що до компонент списку або фрейму можна звертатися також тільки по іменам без вказівки імені фрейму або списку. Це можна зробити за допомогою функції **attach**.

Приклад.

```
> attach(Candidates)
> Algebra-History
[1]  -9  -3 -18 -16  32
```

Ця команда дозволяє до компонент набору даних звертатися тільки по другій частині імені **Algebra, History, Math, Sertificate,** допускаючи, що першої частини імені об'єкту не існує. Зі змінними **Algebra, History, Math, Sertificate** ми можемо працювати як з векторами. Проте відповідний стовпчик у фреймі не змінюється.

Приклад.

```
> Algebra<-2*Algebra
> Algebra
[1] 134 170 148 138 182
> Candidates$Algebra
[1] 67 85 74 69 91
```

Щоб змінити значення компоненти набору даних, потрібно вказати повне ім'я з двох частин змінної (ім’я фрейму$ім’я змінної):

```
>Candidates$Algebra<-Algebra-History
```

Злиття двох фреймів за загальною змінною

Для злиття двох таблиць застосовується функція **merge.** Подивимось, як застосовується ця функція на прикладі.

Приклад. Нам відомі дві таблиці з даними. В першій вказано номер товару і номер категорії товару. В другій – номер товару, ціна з врахуванням знижок, кількість товару. Потрібно об’єднати таблиці по спільному стовпчику.

```
> product.category <- data.frame(product_id = c(1,1,2,2,3),
category_id = c(1,2,1,3,3))
> purchases <- data.frame(product_id = c(1, 2, 3),
+                         totalcents = c(100, 200, 300),
+                         quantity = c(1, 1, 3))
> total<-merge(product.category,purchases)
> total
  product_id category_id totalcents quantity
1          1           1        100        1
2          1           2        100        1
3          2           1        200        1
4          2           3        200        1
5          3           3        300        3
```

Розглянемо інший приклад.

Приклад. Нехай в нас є дві таблиці з оцінками: в першій дані по 5 проблемним студентам, які куратор завантажив з Campus, в другій - дані, які викладачу надав новий лаборант. Потрібно об`єднати ці таблиці.

```
> Sessia_camp<-data.frame(Name=c("Vasylko","Petrenko","Taniy","Mykolin","Fediyko"),"Функан"=c(67,63,59,71,57),"Ймовірність"=c(54,58,74,60,60),"Фізика"=c(78,60,73,68,64),"Філософія"=c(66,88,77,99,77))
> Sessia_lab<-data.frame(Name=c("Vasylko","Petrenko","Taniy","Mykolin","Fediyko","Chpunko","Zabuv"),"Функан"=c(67,63,58,71,57,73,76),"Ймовірність"=c(54,58,74,61,60,77,52),"Фізика"=c(78,60,73,64,64,76,66),"Філософія"=c(66,88,77,69,77,71,70)
> Result<-merge(Sessia_camp,Sessia_lab)
> Result
      Name Функан Ймовірність Фізика Філософія
1  Fediyko     57         60     64       77
2 Petrenko     63         58     60       88
3  Vasylko     67         54     78       66
```

Два набори даних **Sessia_camp, Sessia_lab** зливаються по значеннях вектору **Name**, загального для обох систем даних. Без параметрів **all.x=TRUE** і **all.y=TRUE** були збережені тільки ті спостереження, які мають спільні значення змінної-ключа (якщо не вказати нічого для параметру **by**, то система сама визначить ключ), параметр **sor**t вказує, чи впорядкувати новий набір даних (**sort=TRUE**), чи ні (**sort=FALSE**) за змінною-ключем.

Приклад. Об’єднайте набори даних **Sessia_camp, Sessia_lab** так, щоб були збережені всі спостереження, що належать системі даних **Sessia_camp**.

```
> Result2<-merge(Sessia_camp,Sessia_lab,all.x = T)
> Result2
      Name Функан Ймовірність Фізика Філософія
1  Fediyko     57         60     64       77
2  Mykolin     71         60     68       99
3 Petrenko     63         58     60       88
4    Taniy     59         74     73       77
5  Vasylko     67         54     78       66
```

Приклад. Об’єднайте набори даних, вказавши ключ.

```
> Result3<-merge(Sessia_camp,Sessia_lamb,by="Name")
> Result3
     Name Функан.х Ймовірність.х Фізика.х Філософія.х Функан.у Ймовірність.у
1  Fediyko      57           60      64         77      57           60
2  Mykolin      71           60      68         99      71           61
3 Petrenko      63           58      60         88      63           58
4    Taniy      59           74      73         77      58           74
5  Vasylko      67           54      78         66      67           54
6  Chpunko      NA           NA      NA         NA      73           77
7    Zabuv      NA           NA      NA         NA      76           52
  Фізика.у Філософія.у
1       64         77
2       64         69
3       60         88
4       73         77
5       78         66
6       76         71
7       66         70
```

Дані для деяких змінних відрізнялись, тому стовпчики продублювались з вказанням індексу (х – з першого набору даних, у – з другого), там, де були відсутні дані тепер стоїть символ NA.