Київський національний університет імені Тараса Шевченка
Кафедра теорії ймовірностей, статистики та актуарної математики

Р. Майборода

Самостійна робота по курсу

"Дескриптивна статистика"

Індивідуальні завдання
та рекомендації по виконанню
Робоча версія від 05.09.2017

Київ — 2017

Вступ

Для виконання завдань потрібно встановити R та RStudio на своєму комп'ютері.

Щоб встановити R для Windows зайдіть на сторінку

`http://cran.r-project.org/bin/windows/base/`

і виберіть `Download R 3.4.1 for Windows` (номер версії, скоріше за все, буде вже іншим). Після цього запустіть програму, яка буде завантажена на ваш комп'ютер і відповідайте на її запити.

Якщо вам потрібна версія R для іншої операційної системи, зайдіть на сторінку

`http://www.r-project.org/`

і виберіть там варіант, який вас влаштовує.

Для того, щоб встановити RStudio, зайдіть на сторінку

`www.rstudio.com`

і виберіть там варіант для завантаження. Встановлювати RStudio треба після того, як буде встановлено R.

Книжку [3], присвячену статистичному аналізу даних за допомогою R, можна отримати за адресою:

`http://probability.univ.kiev.ua/userfiles/mre/compsta.pdf`

Завдання 1.

1. Отримайте файл з даними про котирування на американських фондових біржах акцій компаній, що входять до індексу S&P 500. Файл-каталог можна завантажити з сайту компанії Quantquote:

`quantquote.com/files/quantquote_daily_sp500_83986.zip`

Розпакуйте цей архів у зручний для вас каталог (теку). Дані по кожній компанії містяться в в окремому файлі. Імена файлів містять скорочені назви компаній, наприклад, `table_ibm.csv` — файл з даними про котирування акцій компанії IBM. Список компаній з їх скороченими та повними назвами і сферою їх діяльності можна подивитись тут:

`en.wikipedia.org/wiki/List_of_S%26P_500_companies`

Кожен файл містить таблицю у форматі csv з семи стовпчиків:

— дата біржових торгів (формат ррррммдд) — dat

— індикатор — z,

— ціна відкриття — opn,

— максимальне ціна — mx,

— мінімальна ціна — mn,

— ціна закриття — clo,

— обсяг продаж — vol.

2. Знайдіть дані по компаніях, що відповідають вашому варіанту: це компанії, які нумерації за скороченими назвами в алфавітному порядку мають номери від 20N-19 до 20N де N — номер вашого варіанту. (За бажанням, можна вибрати інші компанії, які вас цікавлять і узгодити список з викладачем). Виділіть відповідні файли у окремий каталог. Надалі ви будете працювати тільки з ними.

3. Виберіть з виділених 20 компаній три, найбільш цікаві на ваш погляд. Для цих компаній підрахуйте

— логарифмічні норми прибутку з лагом 1 (log-retutns) за змінною clo (змінна lr);

— логарифми відношення mx/mn (змінна mr)

— логарифми відношення opn/clo (змінна or)

4. Виведіть діаграми розсіювання:

— (mr,or)

— (mr_{-1},lr)

— (or,or)

— (or_{-1},lr)

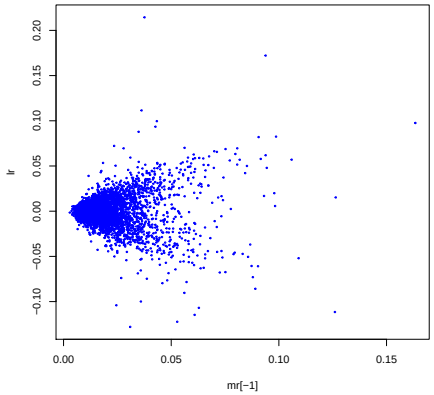

Рис. 1: Діаграма розсівання (mr,lr)

Можна також спробувати вивести діаграми розсіювання для порівняння показників різних фірм.

Найбільш цікаві діаграми опишіть у звіті.

5. Для показників mr, or і lr підрахуйте:

— вибіркові середнє, медіану, середину діапазону.

— дисперсію, середнє квадратичне відхилення, інтерквартильний розмах, ширину діапазону.

Таблицю результатів наведіть у звіті.

Рекомендації по виконанню завдання 1.

Прочитати файл `table_ibm.csv` з каталогу `c:\\rem\\daily` можна, наприклад, так:

```
> setwd("c:\\rem\\daily") # задаємо робочий каталог
> x<-read.csv("table_ibm.csv",header=F) # читаємо таблицю
> # задаємо нові назви змінних-стовпчиків:
> colnames(x)<-c("dat","z","opn","mx","mn","clo","vol")
```

Приклад підрахунку log-returns та логарифму відношення максимальної і мінімальної ціни, відображення діаграми розсіювання:

```
> lr<-log(x$clo[-nrow(x)]/x$clo[-1]) # log-returns з лагом 1
> mr<-log(x$mx/x$mn)
> plot(mr[-1],lr,cex=0.3,col="blue") # діаграма розсіювання (mr,lr)
```

Діаграма розсіювання (mr,lr) зображена на рис. 1. (Бажано пояснити її особливості у звіті, якщо ви їх помітите).

Приклад підрахунку статистик середнього положення:

```
> mean(lr) # вибіркове середнє

[1] -0.0003701923

> median(lr) # медіана

[1] -0.0003594147

> (max(lr)+min(lr))/2 # середина діапазону

[1] 0.04323268
```

Аналогічно можна підрахувати статистики розкиду.

Литература

[1] Карташов М.В. "Імовірність, процеси, статистика". Київ, Видавничо-поліграфічний центр “Київський університет”, 2007, 494 с.

[2] Майборода Р.Є. Регресія: Лінійні моделі.- К. ВПЦ "Київський університет 2007, 296с.

[3] Майборода Р. Комп’ютерна статистика: професійний старт.— 2017

[4] Майборода Р.Є., Сугакова О.В. "Аналіз даних за допомогою пакета R". , 2015 65 с.

[5] Себер Дж. Линейный регрессионный анализ.— М.: Мир, 1980.— 456с.

[6] Турчин В.М. Теорія ймовірностей і математична статистика.- Дніпропетровськ, ІМА-пресс, 2014 - 566 с.

[7] James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R.— Springer NY 2013.— 440p.