Course: Computer statistics

Lecture 1: Descriptive statistics of one-dimensional numerical data

Lecturer: Oleksandr Dykhovychnyi

Лекція 1. Описова статистика одновимірних числових даних

I. Побудова статистичного ряду

II. Найпростіші статистичні оцінки

II.1. Середнє випадкової величини

II.2. Дисперсія випадкової величини

II.3. Середньо квадратичне відхилення

II.4 Медіана й мода вибірки

II.5. Квантилі вибірки

II.6. Коефіцієнт асиметрії і ексцес

II.7. Функція Summary

Первинний аналіз статистичних даних передбачає розрахунок набору числових характеристик, які описують структуру та особливості даних. Сукупність процедур, за допомогою яких обчислюють характеристики даних, називають **описовою** (**дескриптивною**) статистикою, а самі числові характеристики даних, які розраховують, називають **дескриптивними статистиками**.

I. Побудова статистичного ряду

Статистичний ряд за вибіркою формує функція **table(x)**.

Приклади:

```
> x<-c(3.6,7.8,9.6,5.7,8.9,9.6,5.7,8.9,9.6) #вибірка
> x.t<-table(x) # формуємо статистичний ряд
> x.t #друкуємо ряд
```

```
  x
3.6 5.7 7.8 8.9 9.6
 1   2   1   2   3
```

Також для перетворення числової змінної в категоріальну можна скористатися функцією **cut()**:

cut(x, breaks, labels = NULL, include.lowest = FALSE, right = TRUE, dig.lab = 3, ordered_result = FALSE, ...),

де

- **x** — числовий вектор, що перетворюється на фактор;

- **breaks** — або число (> 2), яке задає кількість інтервалів розбиття, або числовий вектор, що задає межі розбиття;

- **labels** — символьний вектор назв категорій;

- **include.lowest** — логічний аргумент , який вказує, чи потрібно включати в інтервал елемент, що співпадає з нижньою (верхньою) границею інтервалу;

- **right** — логічний аргумент, що задає вид інтервалу розбиття: (a, b] (по замовчуванню) або [a, b);

- **dig.lab** — числовий аргумент, що задає число знаків після коми в назві категорій (якщо назви категорій — це інтервали розбиття);

- **ordered_result** — логічний аргумент, що вказує чи потрібно впорядковувати створені категорії.

Приклади. У навчальній групі студенти отримали бали за предмет, але потрібно надати кафедрі дані по символьним оцінкам: F — від 0 до 29, FX — від 30 до 59, E — від 60 до 64, D — від 65 до 74, C — від 75 до 84 , B — від 85 до 94, A — від 95 до 100.

```
> x<-c(83,70,86,51,61,67,80,84,70,64,83,55,88,75,61,70,95,52,75,92,86,89,83,58,51)
> xf<-cut(x,breaks=c(0,29,59,64,74,84,94,100), labels=c("F","FX","E","D","C"
,"B","A"))
> xf
 [1] C  D  B  FX E  D  C  C  D  E  C  FX B  C  E  D  A  FX C  B  B  B
[23] C  FX FX
Levels: F FX E D C B A
> table(xf) #будуємо статистичний ряд (таблицю частот)
xf
 F FX  E  D  C  B  A
 0 5   3  4  7  5  1
Таблиця відносних частот:
```

```
> table(xf)/length(xf) #будуємо таблицю відносних частот

xf
   F   FX    E    D    C    B    A
0.00 0.20 0.12 0.16 0.28 0.20 0.04
```

II. Найпростіші статистичні оцінки параметрів

Нехай $X=(x_1,x_2,...,x_n)$ - вибірка з генеральної сукупності, розподіл якої визначає функція розподілу $F_\xi(x,\theta)$, що залежить від певного параметру θ. Статистичною оцінкою параметра θ ми будемо вважати функцію від вибірки $\theta(X)$, яка у певному сенсі наближає параметр θ. Така оцінка називається **точковою**. Як правило, до точкових оцінок висувають наступні умови:

а) незсуненість означає**,** що математичне сподівання оцінки дорівнює значенню параметра;

б) консистентність – це наближення оцінки за ймовірністю до оцінюваного параметру при збільшенні об'єму вибірки;

в) ефективність – властивість оцінки мати мінімальну оцінку у певному класі оцінок.

Ми не будемо детально розглядати ці поняття. Зосередимось на обчисленні самих оцінок.

1. Середнє випадкової величини - $E\xi$.

Оцінка - в**ибіркове середнє:**

$$\overline{X}=\frac{x_1+x_2+...+x_n}{n}.$$

Реалізує функція ***mean(x, ...),***
де **x -** вектор, матриця або фрейм.

Приклади:

```
>        x<-c(3.6,7.8,9.6,5.7,8.9) #вибірка
>        mean(x) #обчислюємо середнє

[1] 7.12
```

2. Дисперсія випадкової величини-$Var\xi=E\left(\xi-E\xi\right)^2$.

Оцінка - **вибіркова дисперсія:**

$$S_x^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2,$$

або **вибіркова дисперсія виправлена:**

$$S_0^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2.$$

Реалізує функція ***var(x,….),***

де **x** - вектор, матриця або фрейм.

3. Середньо квадратичне відхилення - $\sqrt{Var\xi}$:

Оцінка:

$$S_x = \sqrt{S_x^2}.$$

Реалізує функція ***sd(x,…),***

де **x**- вектор, матриця або фрейм.

Приклади:

```
x<-c(3.6,7.8,9.6,5.7,8.9) #вибірка
> var(x, y, na.rm = FALSE) # обчислюємо дисперсію
[1] 2.9
> sd(x, na.rm = FALSE) # обчислюємо середньо квадратичне відхилення
[1] 2.459065
```

4. Медіана й мода вибірки випадкової величини.

Медіану обчислює функція **median(x,…),**

де **x**- вектор, матриця або фрейм.

Приклади:

```
> x<-c(3.6,7.8,9.6,5.7,8.9,9.6,5.7,8.9,9.6) #вибірка
> median(x) #обчислюємо медіану
[1] 8.9
```

З модою трохи складніше. Потрібно ранжування вибірки. Розберемо на прикладі.

Приклади:

```
> sort(unique(x)) #сортування вибірки
[1] 3.6 7.8 9.6 5.7 8.9

> x.t<-table(x) # формуємо статистичний ряд

> x.t #друкуємо
   x
  3.6  5.7  7.8  8.9  9.6
    1    2    1    2    3
   > sort(unique(x))[which.max(x.t)] #знаходимо моду
  [1] 9.6    мода
```

5. За допомогою функції **quantile(x,…)** обчислюють **квантилі вибірки** заданого розміру α ($0 \le \alpha \le 1$).
Аргументами функції є:

- **x** — числовий вектор (вибірка). Не повинен містити пропущених значень.

- **probs** — числовий вектор (розміри необхідних квантилів).

По замовчуванню обчислюються квартилі.

- **na.rm** — логічний аргумент, який вказує, чи потрібно видаляти з вибірки пропущені значення.

- **names** — логічний аргумент, що вказує, чи потрібно привласнювати імена результатам (вказувати, які квантилі обчислено).

- **type** — додатнє ціле число, що приймає значення від 1 до 9 і визначає алгоритм обчислення квантилів.

Приклади:

```
> quantile(x)
  0%  25%  50%  75% 100%
 3.6  5.7  8.9  9.6  9.6
```

6. Пакет **moments** дозволяє знаходити **коефіцієнт асиметрії** і **ексцес,** які обчислюють функції: **skewness(x,…), kurtosis(x, …)**.

```
> install.packages("moments") # завантаження пакета moments.

>library(moments) #підключення пакета moments.

> set.seed(198)
> x<-rexp(100,5) #моделюємо 100 експоненційно розподілених чисел з
```

#параметром 5

> kurtosis(x, na.rm=TRUE) #обчислення ексцесу

[1] 5.886437
> skewness(x, na.rm = TRUE) # обчислення коефіцієнта асиметрії

[1] 1.665976

7. У системі **R** є можливість і набагато швидше оцінити основні параметри вибірки . Це можна зробити за допомогою функції загального призначення - **summary(x)**. Застосуємо до тієї ж самої вибірки

Приклади:

>summary(x)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.600	5.700	8.900	7.711	9.600	9.600

Всього однієї команди достатньо для знаходження мінімального (**Min**) і максимального (**Max**) значень вибірки , медіани (**Median**), вибіркового середнього (**Mean**), першого (**1st Qu**) і третього (**3rd Qu**) квартілів.

Подібні обчислення можна отримати для всього фрейму даних.
Приклади:

>weight # набір містить вагу трьох груп овочів

	weight1	weight2	weight3
1	1.6	1.7	1.90
2	1.9	1.19	1.16
3	1.29	1.2	0.80
4	1.5	2.1	1.15
5	2.7	2.9	0.90
6	1.5	1.6	1.60

>summary(weight) summary по всіх параметрах

```
     Weight1         weight2          weight3
Min.   :1.290   Min.   :1.190   Min.   :0.8000
1st Qu.:1.500   1st Qu.:1.300   1st Qu.:0.9625
Median :1.550   Median :1.650   Median :1.1550
Mean   :1.750   Mean   :1.783   Mean   :1.2517
3rd Qu.:1.825   3rd Qu.:2.000   3rd Qu.:1.4900
Max.   :2.700   Max.   :2.900   Max.   :1.9000
```

Часто для розрахунку однакової характеристики для матриці використовується функція **apply()**. В наступному прикладі формуємо матрицю концентрацій певної речовини, яка вимірюється у фіксовані п’ять моментів часу, і знайдемо середньо квадратичне відхилення по кожному рядку.

Приклади:

```
> d15=c(0.005,0.008,0.010,0.005)
> d16=c(0.004,0.005,0.015,0.008)
> d17=c(0.004,0.010,0.012,0.009)
> d18=c(NA,NA,NA,NA)
> d19=c(0.008,0.011,0.014,0.015)
> d20=c(0.009,0.011,0.014,0.007)
> d21=c(0.007,0.009,NA,NA)
> #Формуємо матрицю спостережень :
> fm=rbind(d15,d16,d17,d18,d19,d20,d21)
> apply(fm,1,sd,na.rm=T) # обчислюємо СКВ з урахуванням пропущених
                         #значень
```

```
      d15         d16         d17       d18         d19         d20         d21
0.002449490 0.004966555 0.003403430        NA 0.003162278 0.002986079 0.001414214
```

Функції для обчислення описових статистик зібрано у таблиці 1.

Таблиця 1

Статистика	Функція
Вибіркове середнє	mean(x)
Геометричне середнє	prod(x)~(l/length(x))
Гармонійне середнє	l/mean(l/x)
Зрізане середнє	mean(x,trim=a)
Медіана	median(x)
Виправлена вибіркова дисперсія	var(x)
Середньоквадратичне відхилення	sd(x)
Середнє абсолютне відхилення	mean(abs (x-mean(x)))
Інтерквартильний розмах	IQR(x)