Самостійна робота з дескриптивної статистики

Горбунов Даніел Денисович,
III курс бакалаврату,
комп'ютерна статистика

Варіант №4

8 вересня 2019 р.

1 Завдання №1

1.1 Дані

Для міст України: Київ, Дніпро, Одеса, Львів, Харків, Івано-Франківськ, Житомир:

1. *Перший набір: ціни на абонемент на 1 місяць у фітнес-клубі на одну особу;*
2. *Другий набір: ціни на 1 кілограм помідорів.*

Для кожного набору обчисліть такі дескриптивні статистики вибірки:

- *Вибіркове середнє;*
- *Середнє геометричне;*
- *Середнє гармонійне;*
- *Медіана;*
- *Середина діапазону;*
- *Дисперсія;*
- *Середньоквадратичне відхилення;*
- *Інтерквартильний розмах;*
- *Ширина діапазону;*
- *Коефіцієнт варіації.*

Дані з файлів **countries_fitness.csv** та **countries_tomatoes.csv**.

Names	Values
Kiev	526.96
Dnipro	388.81
Odesa	559.75
Lviv	458.16
Kharkiv	489.5
Ivano-Frankivsk	420
Zhytomyr	351

Names	Values
Kiev	39.78
Dnipro	38.08
Odesa	35.89
Lviv	34.1
Kharkiv	24.33
Ivano-Frankivsk	32.5
Zhytomyr	27.05

1.2 Формули обчислення та їх реалізація в R

Нехай $X = \{X_1, ..., X_n\}$ - певна вибірка, де:
X_j - значення досліджуваної змінної у j-тому спостереженні,
n - кількість елементів у вибірці.
Варіаційний ряд має наступний вигляд: $\min\limits_{1 \leqslant j \leqslant n} X_j = X_{[1]} \leqslant X_{[2]} \leqslant ... \leqslant X_{[n]} = \max\limits_{1 \leqslant j \leqslant n} X_j$

Вибіркове середнє для вибірки X визначається за формулою:

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j \tag{1}$$

Середнє геометричне, що визначається для таких вибірок X, у яких значення змінної X_j приймають лише додатні значення:

$$GM(X) = \sqrt[n]{\prod_{j=1}^{n} X_j} \tag{2}$$

Середнє гармонійне дорівнює наступному:

$$HM(X) = \frac{n}{\sum_{j=1}^{n} \frac{1}{X_j}} \tag{3}$$

Вибіркова медіана для вибірки X обчислюється за формулою:

$$med(X) = \begin{cases} X_{[(n+1)/2]}, & \text{якщо n - непарне;} \\ \frac{1}{2}(X_{[n/2]} + X_{[n/2+1]}), & \text{якщо n - парне.} \end{cases} \tag{4}$$

Середина діапазону:

$$MR(X) = \frac{1}{2}(X_{[1]} + X_{[n]}) \tag{5}$$

Будемо застосовувати формулу для обчислення **виправленої вибіркової дисперсії**. Відрізняється від звичайної нормуючим множником $(n-1)/n$:

$$S_0^2(X) = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2 \tag{6}$$

Середньоквадратичне відхилення - квадратний корінь від значення вибіркової дисперсії:

$$S_0(X) = \sqrt{S_0^2(X)} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2} \tag{7}$$

Інтерквартильний розмах:

$$IQ(X) = Q_3(X) - Q_1(X), \tag{8}$$

де $Q_3(X)$ та $Q_1(X)$ визначаються наступним чином:

$$Q_2(X) = med(X)$$

$$Q_1(X) = med(\{X_{[j]} \in X | \min_{1 \leqslant j \leqslant n} X_j = X_{[1]} \leqslant X_{[2]} \leqslant ... \leqslant X_{[n]} = Q_2(X)\})$$

$$Q_3(X) = med(\{X_{[j]} \in X | Q_2(X) = X_{[1]} \leqslant X_{[2]} \leqslant ... \leqslant X_{[n]} = \max_{1 \leqslant j \leqslant n} X_j\})$$

Ширина діапазону:

$$Range(X) = X_{[n]} - X_{[1]} \tag{9}$$

Коефіцієнт варіації:

$$CV(X) = \frac{S_0(X)}{\overline{X}} \tag{10}$$

Маючи всі необхідні формули для обчислення статистик, спробуємо реалізувати кожну в R.

1. Вибіркове середнє:

```
sample_mean <- function(sample)
{
  result <- mean(sample)
  result
}
```

2. Середнє геометричне:

```
geometric_mean <- function(sample)
{
  result <- prod(sample) ^ (1/length(sample))
  result
}
```

3. Середнє гармонійне:

```
harmonic_mean <- function(sample)
{
  result <- length(sample)/sum(1/sample)
  result
}
```

4. Вибіркова медіана:

```
sample_median <- function(sorted_sample)
{
  result <- median(sorted_sample)
  result
}
```

5. Середина діапазону:

```
mid_range <- function(sample)
{
  result <- 0.5 * (max(sample) + min(sample))
}
```

6. Вибіркова дисперсія:

```
variance <- function(sample)
{
  result <- var(sample)
  result
}
```

7. Середньоквадратичне відхилення:

```
standard_deviation <- function(sample)
{
  result <- sd(sample)
  result
}
```

8. Інтерквартильний розмах:

```
iq <- function(sorted_sample)
{
  q2 <- sample_median(sorted_sample)
  q1 <- sample_median(sorted_sample[sorted_sample <= q2])
  q3 <- sample_median(sorted_sample[sorted_sample >= q2])
  result <- q3 - q1
  result
}
```

9. Ширина діапазону:

```
sample_range <- function(sample)
{
  result <- max(sample) - min(sample)
  result
}
```

10. Коефіцієнт варіації:

```
cv <- function(sample)
{
  result <- standard_deviation(sample)/sample_mean(sample)
  result
}
```

1.3 Виконання завдання

Опишемо головну функцію для виконання умов завдання та збереження результатів у вигляді таблиці:

```
examine <- function(sample, filename)
{
  sorted_sample <- sort(sample)

  result_df <- data.frame(
    sample_mean        (sorted_sample),
    geometric_mean     (sorted_sample),
    harmonic_mean      (sorted_sample),
    sample_median      (sorted_sample),
    mid_range          (sorted_sample),
    variance           (sorted_sample),
    standard_deviation(sorted_sample),
    iq                 (sorted_sample),
    sample_range       (sorted_sample),
    cv                 (sorted_sample)
  )

  names(result_df) <- c(
      "mean", "geometric mean", "harmonic mean",
      "median", "mid-range", "variance",
      "standard deviation", "iq", "range", "cv"
  )

  write.csv(
    result_df,
    file=filename
  )
}
```

Проведення операцій з наборами даних в кінці програми:

```
fitness_data <- read.csv(
    file="/home/fourier-transform/R/r_proj/countries_fitness.csv",
  header=TRUE, sep=";"
)

tomato_data <- read.csv(
  file="/home/fourier-transform/R/r_proj/countries_tomatoes.csv",
  header=TRUE, sep=";"
)

examine(fitness_data[,2], "/home/fourier-transform/R/r_proj/result_frame_1.csv")
examine(tomato_data[,2], "/home/fourier-transform/R/r_proj/result_frame_2.csv")
```

Після виконання вищенаведених інструкцій програми отримаємо дві таблиці з певними даними.

result_frame_1.csv, в якому зберігаються результати обчислень після обробки набору даних файлу **countries_fitness.csv**:

mean	geometric mean	harmonic mean	median	mid-range
456.311428571429	450.9406123977	445.529580368722	458.16	455.375

variance	standard deviation	iq	range	cv
5626.85281428571	75.0123510782439	103.825	208.75	0.164388499567246

result_frame_2.csv, в якому зберігаються результати обчислень після обробки набору даних файлу **countries_tomatoes.csv**:

mean	geometric mean	harmonic mean	median	mid-range
33.1042857142857	32.6618159997933	32.1947846244488	34.1	32.055

variance	standard deviation	iq	range	cv
32.0136952380952	5.65806461946974	7.21	15.45	0.170916378269055

Зауваження: Очевидно, при обчисленні вибіркової дисперсії за формулою:

$$S^2(X) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2,$$

отримали б інші значення для *variance*, *standard deviation* та *cv*.

1.4 Обчислення статистик без використання можливостей комп'ютера

Виконані в зошиті..

1.5 Висновок

Мова програмування R - це універсальний інструмент для швидких обчислень.
Якщо мовити про результати аналізу вищенаведених наборів даних, то можна дати наступну характеристику: розкид першої вибірки майже еквівалентний другому. Більшість значень, отриманих за допомогою обчислень без використання комп'ютера, співпали з результатами роботи програми.