УДК 658.512.22.011.56

РЕАЛИЗАЦИЯ ВСПОМОГАТЕЛЬНОЙ ДИАГРАММЫ ВИЗУАЛИЗАЦИИ ОТСЕВА ВЫБРОСОВ БОЛЬШОЙ ВЫБОРКИ

Напрасников В.В. [1], **Напрасникова Ю.В.** [2], **Соловьев А Н.** [3]
Белорусский национальный технический университет,
Минск, Республика Беларусь;
2) Унитарное предприятие «Калийпроект»
Минск, Республика Беларусь;
3) Крымский инженерно-педагогический университет
Симферополь, Российская Федерация.

Целью данной работы является программная реализация построения боксплота («ящик с усами») для наглядного представления о распределении значений в среде MATHCAD 15. Boxplot («ящик с усами») это графический и аналитический метод описательной статистики, предназначенный для визуализации распределения числовых данных, оценки их локации, разброса, асимметрии и идентификации потенциальных выбросов в выборке.

Особенностью представляемой программной реализации является добавление отсечки, соответствующей уровню экстремальных (жестких) выбросов, что повышает наглядность представления обрабатываемой информации и повышает достоверность принимаемых решений.

Пусть имеется одномерная выборка данных $X = (x_1, x_2, \dots, x_n)$, упорядоченная по возрастанию: $x_{1p} \le x_{2p} \le \cdots \le x_{np}$.

Модель основывается на пяти статистиках:

Q_0 (Минимум): x_{1p} (наименьшее наблюдение, без учета выбросов).

Q_1 (Первый квартиль, 25-й перцентиль). Для выборки объёмом n позиция определяется как:

$$L_1 = 0.25 \cdot (n + 1). \quad (1)$$

Q_2 (Медиана, 50-й перцентиль): Значение, делящее выборку пополам.

$$L_2 = 0.5 \cdot (n + 1). \quad (2)$$

Q_3 (Третий квартиль, 75-й перцентиль). Значение, ниже которого расположено 75% данных.

$$L_3 = 0.75 \cdot (n + 1). \quad (3)$$

Q_4 (Максимум): x_{np} (наибольшее наблюдение, без учета выбросов).

Определение межквартильного размаха (IQR — Interquartile Range):

$$IQR = Q_3 - Q_1. \quad (4)$$

Это робастная мера статистического разброса данных, нечувствительная к экстремальным значениям.

Выбросы (Outliers):

Все наблюдения x_i, такие что $x_i < Q_1 - k \cdot IQR$ или $x_i > Q_3 + k \cdot IQR$, идентифицируется как потенциальные выбросы.

Для идентификации экстремальных (жестких) выбросов используется коэффициент $k = 3.0$. Соответствующим образом строятся дополнительные усы (отсечки) Lower Whisker2, L2), (Upper Whisker2, U2).

Фрагмент документа с процедурой отсева жестких выбросов представлен на рисунке 1.

```
USI_OTSEV_JESTK(X) := | Nrows ← rows(X)
                      | Xsr ← mean(X)
                      | S_X ← stdev(X)
                      | KV_1 ← KVARTIL_N(X, 1)
                      | KV_3 ← KVARTIL_N(X, 3)
                      | R_μ ← KV_3 − KV_1
                      | for J ∈ 1 .. Nrows
                      |   Num_(J−1) ← J
                      | NumX ← augment(Num, X)
                      | NumX_rez ← NumX
                      | K ← −1
                      | for J ∈ 0 .. Nrows − 1
                      |   if ⌈(X_J < KV_1 − 3·R_μ) ∨ (X_J > KV_3 + 3·R_μ)⌉
                      |     | K ← K + 1
                      |     | Nrows_del_(K, 0) ← J + 1
                      |     | Nrows_del_(K, 1) ← X_J
                      |     | NumX_rez ← Del_ROWS_N(NumX_rez, J − K)
                      | return (Nrows_del  NumX_rez)
```

Рисунок 1. Фрагмент документа с процедурой отсева жестких выбросов

Использование инструмента для решения реальной задачи по 150 геологическим пробам представлено на рисунке 1 и 2. Найдено 9 мягких и 10 экстремальных выбросов, фрагмент документа с результатами вычислений представлен на рисунке 2.

ТОЛЬКО для МЯГКИХ выбросов НОМЕРА точек и значения точек удаляемых

Rez_MIAGK := USI_OTSEV_MIAGK(X)

$$\left(Rez\_MIAGK^{T}\right)_{0}^{T} = \begin{pmatrix} 11 & 53 & 66 & 68 & 103 & 105 & 126 & 131 & 142 \\ 2.114 & 2.212 & 2.211 & 2.208 & 2.114 & 2.115 & 2.217 & 2.228 & 2.101 \end{pmatrix}$$

ТОЛЬКО для ЖЕСТКИХ выбросов НОМЕРА точек и значения точек удаляемых

Rez_JEST := USI_OTSEV_JESTK(X)

$\left(Rez\_JEST^{T}\right)_{0}^{T} =$

	0	1	2	3	4	5	6	7	8	9	10
0	10	12	24	27	85	87	106	107	120	128	141
1	2.093	2.05	2.094	2.066	2.082	2.056	2.087	2.072	2.289	2.262	2.092

Рисунок 2. Результаты расчета

На рисунке 3 представлена соответствующая диаграмма.

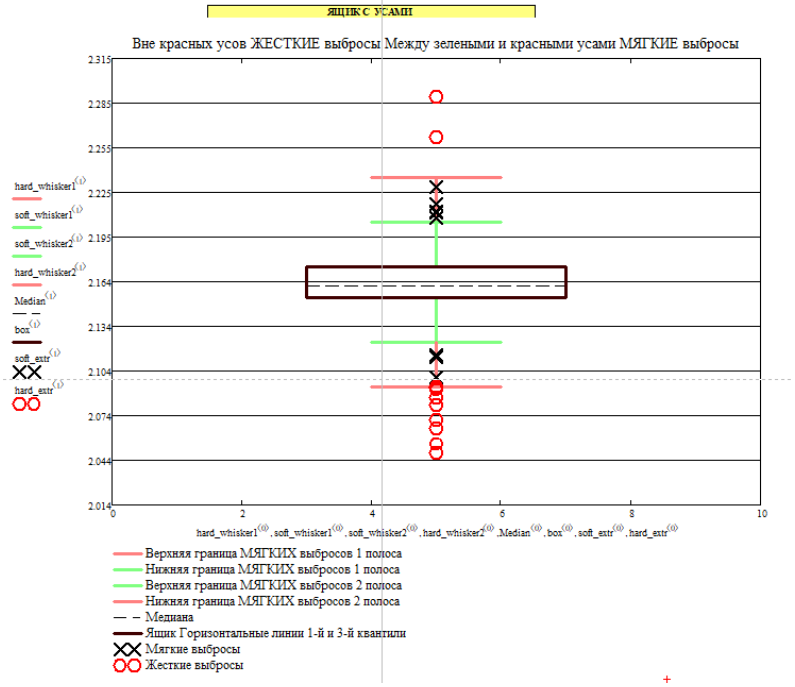

Рисунок 3. Диаграмма с результатами.

Вывод.

В среде MATHCAD создан инструментарий, позволяющий в наглядном виде отображать количество и расположение мягких и экстремальных (жестких) выбросов в выборке. Это позволяет исследователю обоснованно принимать решение об исключении соответствующих наблюдений из выборки.

Созданная диаграмма является более наглядной по сравнению с подобной диаграммой в среде EXEL и позволяет решать поставленную задачу без привлечения другой программной среды.

1. Tukey, J.W. Exploratory Data Analysis [Text] / J.W. Tukey. -Addison-Wesley-Reading, MA, 1977.