УДК 004.02

Ерофеев Ермак Владимирович,
студент, Ангарский государственный технический университет,
e-mail: ermak.080@mail.ru
Кулакова Ирина Михайловна,
к.т.н., доцент, Ангарский государственный технический университет,
e-mail: iyelkina@mail.ru

РАЗРАБОТКА ИНСТРУМЕНТА ВИЗУАЛИЗАЦИИ СТАТИСТИЧЕСКИХ ДАННЫХ

Erofeev E.V., Kulakova I.M.

DEVELOPMENT OF A TOOL FOR VISUALIZING STATISTICAL DATA

Аннотация. В работе произведен анализ возможностей инструмента «горизонтальная диаграмма размаха» для визуализации статистических данных, разработан алгоритм построения инструмента, обладающего заданными свойствами, осуществлена реализация данного алгоритма.

Ключевые слова: визуализация данных, диаграмма размаха, статистические графики.

Abstract. The article analyzes possibilities of the tool horizontal swing diagram for visualization of statistical data, develops an algorithm of construction of a tool with specified properties, implements this algorithm.

Keywords: data visualization, scale diagram, statistical graphs.

В век информационных технологий и больших данных всё чаще возникает потребность в их быстром анализе и представлении в понятной для человека форме. Визуализация – это один из способов работы с данными, который предоставляет возможность быстро их проанализировать. Она позволяет рассмотреть данные с разных сторон, донести до пользователя необходимую информацию кратчайшим путём, скрывая за визуальным оформлением большие объемы данных, что в последствии может избавить от работы с ненужной информацией и помочь при принятии правильных решений. Но не всякая визуализация может быть использована для работы с данными так чтобы её можно было назвать качественной. Качественная визуализация должна характеризоваться несколькими факторами [1]:

– Оригинальность. Хорошая визуализация данных должна формировать нестандартный взгляд на вещи, выводить понимание данных на новый уровень;

– Информативность. Важная цель визуализации донести до пользователя нужную информацию, она не должна быть перегружена большими объемами данных, отображение данных должно быть кратким и понятным каждому;

– Аутентичность. Использование типовых элементов визуализации, таких как оси, легенды данных, заголовки диаграммы.

Одним из качественных методов для визуализации большого объёма данных является диаграмма размаха (boxplot), которая представляет стандар-

тизированный способ графического представления распределения набора данных на основе пяти статистических показателей, таких как:

- минимум – значение диаграммы, которое вычисляется, как Q1-1,5*IQR;
- первый квартиль [Q1] – показывает то, что 25% данных находится ниже этого значения;
- медиана – представлена вертикальной линией внутри прямоугольника, является значением, которое разбивает данные на две половины, при этом 50% значений будут ниже, а 50% выше него;
- третий квартиль [Q3] – показывает то, что 75% данных находится ниже этого значения;
- максимум – значение диаграммы, которое вычисляется, как Q3+1,5*IQR.

Диаграмма размаха отражает важные статистические данные в форме коробчатой конструкции с границами, а также прямыми исходящими из противоположных сторон коробки, если выбросов нет, то они простираются до минимального и максимального значения. Такое краткое и качественное представление данных позволяет сразу оценить основную направленность, изменчивость и распространение набора данных без необходимости проведения сложных статистических расчетов и наличие потенциальных выбросов в данных. Пример типичной диаграммы размаха представлен на рисунке 1 [2, 3].

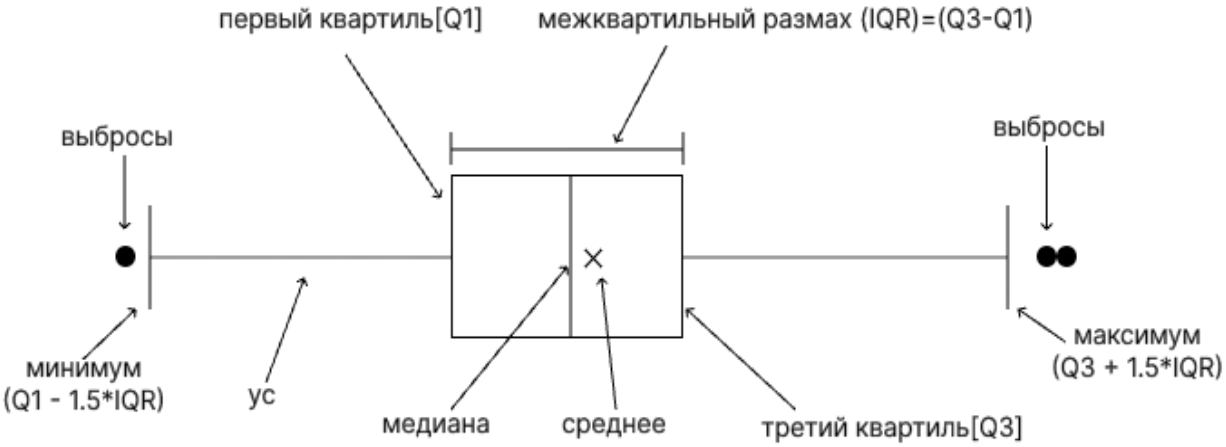

Рисунок 1 – Описание диаграммы размаха (boxplot)

Для демонстрации возможностей диаграммы размаха проводились три независимых теста (Таблица 1). В качестве тестовых данных использовались выборки, обладающие различными статистическими свойствами.

Таблица 1

Тестовые данные

Data-1	14, 16, 17, 20, 11, 13, 17, 16, 22, 15, 13, 13
Data-2	11, 38, 15, 35, 11, 8, 32, 10, 16, 16, 27, 18
Data-3	0,030, 0,321, 0,32, 0,003, 0,2, 0,21, 32, 10, 16, 16, 27, 18

Для приведенных наборов данных были определены характеристики, необходимые для построения диаграммы размаха (Таблица 2). По данным таблицы 2 построены диаграммы (рисунок 2).

Таблица 2

Статистические характеристики тестовых данных

Минимум (min)	Первый квартиль (Q1)	Медиана(M)	Третий квартиль (Q3)	Максимум (max)
11	13	15	17	22
8	11	16	28,25	38
0,003	0,205	5,1605	16,5	32

Такое краткое и качественное представление данных позволяет сразу оценить основную направленность, изменчивость, распространение набора данных и наличие потенциальных выбросов. По ширине ящика видно, что для набора Data-1 характерен наименьший разброс данных относительно средних значений, а набор Data-3 имеет левостороннюю асимметрию за счёт сдвига коробки влево. Относительные положения среднего квадратического отклонения и медианы говорит о наличии выбросов (Data-2) – если они не совпадают или их отсутствии (Data 1). Таким образом, наглядность этого инструмента визуализации не оставляет сомнений.

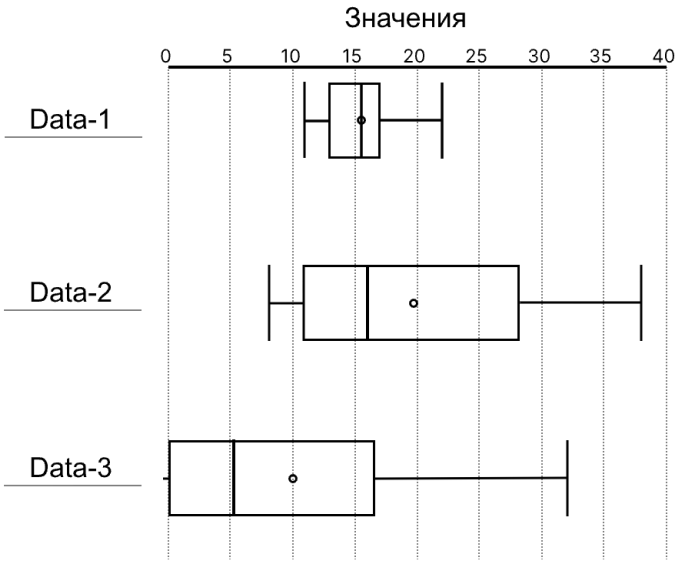

Рисунок 2 – Диаграммы размаха (boxplots)

Однако его использование незаслуженно не получило широкого распространения и лишь немногие ресурсы содержат готовый инструментарий для его отрисовки.

Поэтому существует необходимость разработки алгоритма построения данного графика и его программной реализации. Методика построения рассмотрена на примере графического компонента TCanvas в среде Delphi 12 [4].

Входными данными алгоритма построения диаграммы являлись:

1. Массив чисел DataSet, представляющий выборку статистических данных для визуализации, предварительно отсортированный по возрастанию.
2. Размеры области построения высота (height) и ширина (width), заданные в пикселях.

Определение геометрических параметров графика осуществляется только исходя из размеров области построения. Для однозначной идентификации элементов графика по исходным данным определяются следующие характеристики, зависящие от значений height и width. Ниже представлены некоторые этапы построения диаграммы, их отображение проиллюстрировано рисунком 3.

При описании алгоритма используются команды:

moveTo(X,Y) – установка текущего положения указателя.

lineTo(X,Y) – вычерчивание линии от текущей позиции до указанной точки.

- Расположение оси.

Отображается по центру горизонтали области построения:

moveTo(3, height / 2);

lineTo(width – 3, height / 2).

- Масштаб оси.

Зависит от экстремальных значений массива DataSet определяется как:

scaleValue = (max – min) / width

- Положение и размер ящика.

Определение положения и размера ящика требует двух значений, рассчитанных по выборке: Q1 и Q3 для определения координаты по оси x и высоты, которая вычисляется в процентном отношении от height, размер этого отношения задан параметром attitude = 0,8.

```
moveTo(scaleValue*Q1, height*((1–attitude)/2));
lineTo(scaleValue * Q1, height*attitude);
lineTo(scaleValue * Q3, height*attitude);
lineTo(scaleValue * Q3, height*((1–attitude)/2));
lineTo(scaleValue * Q1, height*((1–attitude)/2));
```

- Положение медианы.

Положение медианы задаётся в форме вертикальной линии по высоте ящика:

```
moveTo(scaleValue*M, height * ((1– attitude) /2));
lineTo(scaleValue * M, height*attitude).
```

- Длина усов и вертикальные линии на концах.

moveTo(scaleValue*Q1, height/2));
lineTo(scaleValue * Q1-1.5*(Q3-Q1), height/2);
lineTo(scaleValue * Q1-1.5*(Q3-Q1), height*((1–attitude)/2));
lineTo(scaleValue * Q1-1.5*(Q3-Q1), height*attitude);
moveTo(scaleValue*Q1, height/2));
lineTo(scaleValue * Q3+1.5*(Q3-Q1), height/2);
lineTo(scaleValue * Q3+1.5*(Q3-Q1), height*((1–attitude)/2));
lineTo(scaleValue * Q3+1.5*(Q3-Q1), height*attitude).

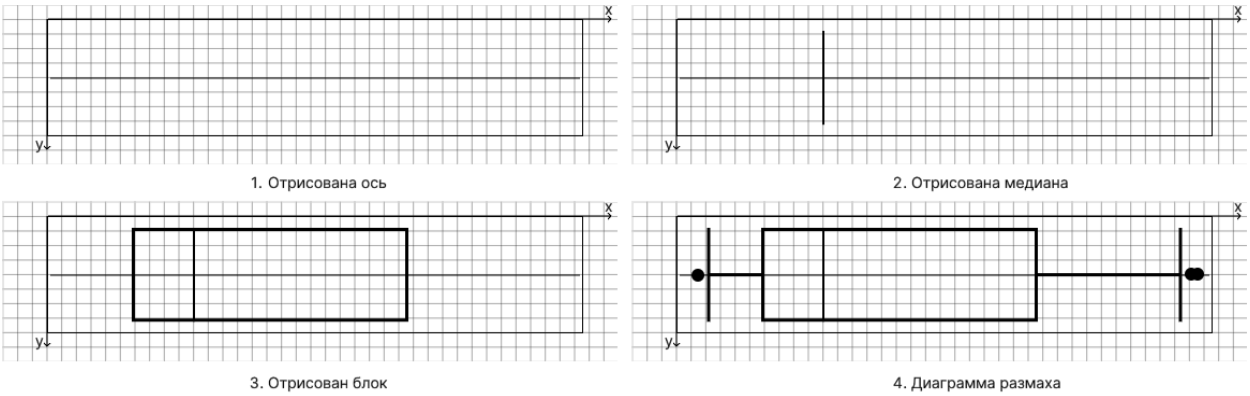

Рисунок 3 – Этапы отрисовки диаграммы размаха (boxplots)

Таким образом, использование данного подхода позволит строить диаграммы размаха на любых данных, автоматически масштабировать при этом размеры элементов, адаптируя их под размеры области построения. Также этот алгоритм может быть использован с любым инструментарием для разработки графических интерфейсов (GUI – graphical user interface), обладающих сходными свойствами и методами.

ЛИТЕРАТУРА

1. **Харрис Роберт Л.** Information Graphics: A Comprehensive Illustrated Reference / Роберт Л. Харрис – М.: Oxford University Press, 2000 – 448 с.

2. **Хафф Даррелл.** How to Lie with Statistics / Даррелл Хафф– М.: WW Norton & Company, 1993 144 с.

3. **Клинтон А., Шевляков Г.Л.** Обнаружение выбросов с помощью боксплотов, основанных на новых высокоэффективных робастных оценках масштаба // Информатика, телекоммуникации и управление – 2013. [Электронный ресурс]. – URL: https://cyberleninka.ru/article/n/obnaruzhenie-vybrosov-s-pomoschyu-boksplotov-osnovannyh-na-novyh-vysokoeffektivnyh-robastnyh-otsenkah-masshtaba (01.05.2024).

4. **Бунаков, П.Ю.** Практикум по решению задач на ЭВМ в среде Delphi : учебное пособие / П. Ю. Бунаков, А. К. Лопатин. — Москва : ФОРУМ : ИНФРА-М, 2019. — 304 с.