УДК 004.891.3 https://doi.org/10.26160/2541-8637-2023-11-198-201

АНАЛИЗ МАССИВА МЕДИЦИНСКИХ ДАННЫХ НА НАЛИЧИЕ ВЫБРОСОВ

Серобабов А.С.[1], Серобабова А.Л.[2]
[1]Омский государственный технический университет, Омск;
[2]Московский областной филиал Московского университета Министерства внутренних дел Российской Федерации имени В.Я. Кикотя, пос. Старотеряево

Ключевые слова: выброс, квартиль, интерквартильный размах, диаграмма размаха.
Аннотация. В работе рассмотрены вопросы анализа и нахождения выбросов в медицинских данных. Предложено использовать диаграммы размахы для наглядного, компактного и информативного представления медицинских данных. По полученным результатам сделаны выводы о необходимости совместного использования биологических факторов при анализе результатов о выбросах. Полученные результаты позволяют обосновать включение предполагаемых выбросов в основную выборку данных.

ANALYSIS OF MEDICAL DATA FOR OUTPUTS

Serobabov A.S.[1], Serobabova A.L.[2]
[1]Omsk state technical university, Omsk;
[2]Moscow Regional Branch of the Moscow University of the Ministry of Internal Affairs of Russia named after V.Ya. Kikot, Staroteryaevo village

Keywords: outlier, quartile, interquartile range, box-plot.
Abstract. The paper looks at analysing and finding outliers in medical data. The use of box-plot to present medical data in a clear, compact and informative way is suggested. The results are used to conclude that biological factors should be combined in the analysis of the emission results. The results justify the inclusion of expected emissions in the main data sample.

Введение. В настоящее время в медицинской деятельности большое внимание уделяется возможности сбора и хранения больших объемов информации о пациенте для ее последующего анализа и использования при постановке диагноза. Выбросы в данных (результат измерения, который веделяется из общей выборки) могут значительно исказить результаты обследования пациента и привести к постановке ошибочного диагноза. Для обработки медицинской информации, в том числе имеющихся выбросов, применяются методы первичного анализа данных и их подготовки для дальнейшего изучения.

В исследовании предложено при обработке медицинской информации опираться на визуальное представление данных в виде диаграмм размаха, позволяющих определять выбросы посредством графического отображения всей совокупности исследуемых данных.

Результаты исследований и обсуждение

В качестве исходной информации в исследовании использованы данные 149 пациентов, проходивших диспансеризацию в медицинских учреждениях города Омска [1]. Каждый пациент характеризуется параметрами P_h (рост

пациента), P_w (ширина желчного пузыря), P_a (возраст пациента), L_{obr} (содержание рецепторов, воспринимающих лептин в крови), L_{lep} (содержание лептина в крови).

Для наглядного отображения данных и поиска в них выбросов использована диаграмма размаха (ящик с усами), предложенная Дж. Тьюки [2]. Диаграмма наглядно представляет большое количество информации в компактной форме, описывая расределение одномерных данных и предоставляя информацию о параметрах положения, масштаба, асимметрии.

Диаграмма размаха состоит из элементов «тело» и «усы». «Тело» показывает интерквантильный размах распределений (25% и 75% перцентили (значение, которое исследуемая величина не превышает с фиксированной вероятностью)). «Усами» отображают весь разброс точек, кроме выбросов, к которым относятся значения за границами «усов», соответствующих формулам:

$$X_1 = Q_1 - k \cdot (IQR); \quad (1)$$

$$X_2 = Q_3 - k \cdot (IQR); \quad (2)$$

где X_1 – нижняя граница уса; X_2 – верхняя граница уса; Q_1 – первый квартиль (содержит 0,25 часть с минимальными значениями выборки); Q_3 – третий квартиль (содержит 0,25 часть с максимальными значениями выборки); k – коэффициент, наиболее часто употребляемое значение, установлено эмпирически и равно 1,5 [3]; $IQR = (Q_1 - Q_3)$ – интерквартильный размах.

Медиана обозначается толстой линией внутри тела и делит интерквартильную область на две части. Если медиана не находится на равном расстоянии от границ тела, то данные асимметричны.

Для определения выбросов использовано среднеквадратическое отклонение, которое характеризует величину отклонений значений j-го параметра от среднего:

$$\sigma\sqrt{D(K_j)}, \quad (3)$$

где $D(K_j)$ – дисперсия j-го параметра.

На рисунке 1 представлена диаграмма размаха, построенная по данным пациентов для параметров P_h, P_w, P_a, L_{obr}, L_{lep}. Показано, что оцениваемые параметры P_a, L_{obr}, P_p не имеют четко выраженных выбросов. Однако выделенные области у параметров L_{lep} и P_p определяются как выбросы, поскольку превышают значение отклонения в 2,698σ.

Учитывая специфику заболевания, проявляющуюся у людей с ожирением, результаты «выброса 1» не стоит исключать из общей выборки, а необходимо принять во внимание и при постановке диагноза учитывать принадлежность пациента к выявленному выбросу.

На рисунке 2 представлена диаграмма размаха параметра L_{lep} в зависимости от половой принадлежности пациентов.

Как можно заметить, значения медианы относительно «тела» не смещены, но относительно друг друга отличаются в 1,6 раза. Такой результат объясняется биологическим фактором [4, 5], т.е. разным средним количеством лептина у мужчин и женщин.

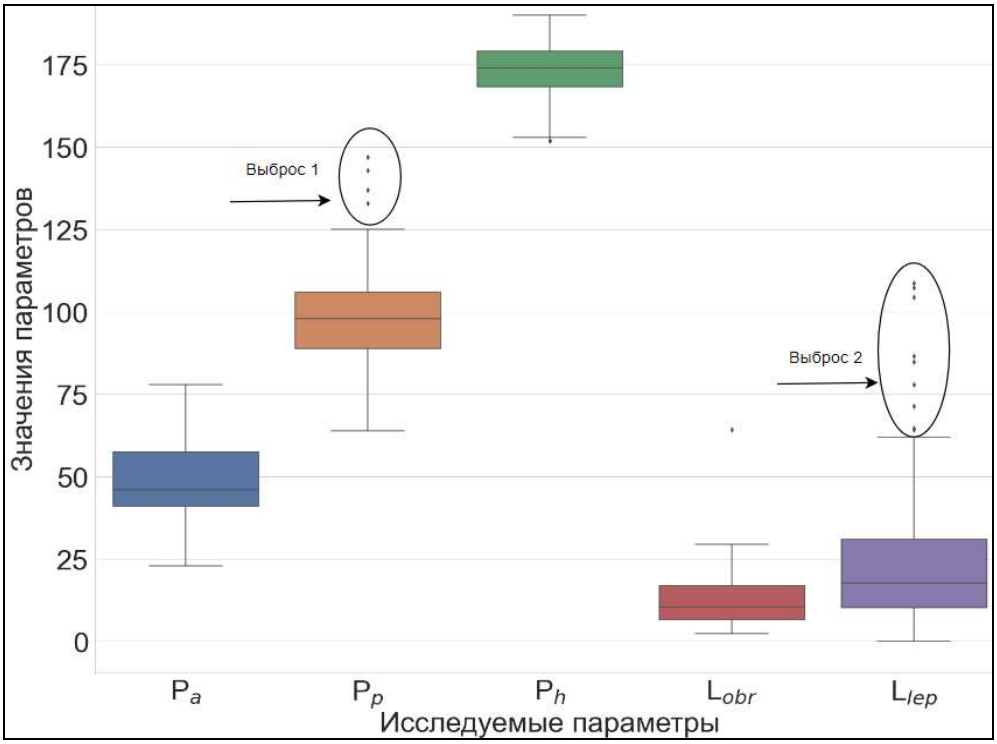

Рис. 1. Диаграмма размаха исследуемых параметров пациента: L_{obr}, P_p, P_a, L_{lep}, P_h

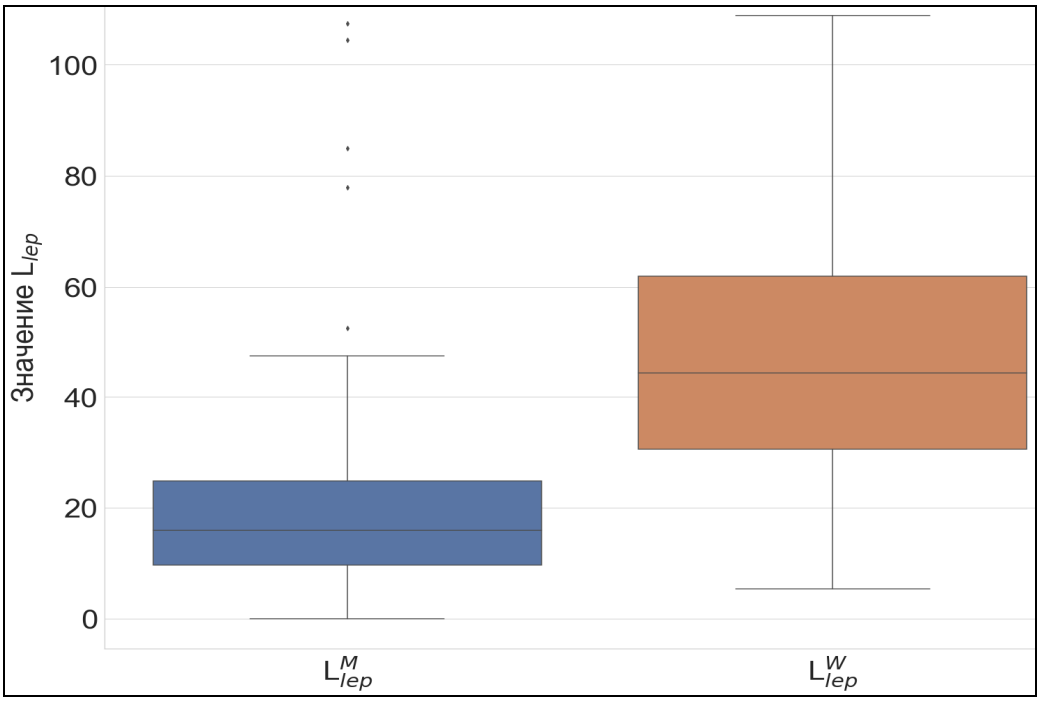

Рис. 2. Диаграмма размаха параметра L_{lep} в зависимости от пола: L^M_{lep} – распределение параметра для мужчин, L^W_{lep} – распределение параметра для женщин

Также прослеживается связь принадлежности некоторых пациентов к двум выборкам сразу, что позволяет сделать предположение о том, что существует зависимость между параметрами лабораторных значений и физиологического показателя P_p (вес).

Заключение. В результате проведенного исследования получены обоснования имеющихся выбросов в медицинских данных. Предложено при анализе медицинских данных учитывать биологические факторы конкретного

параметра. Представляется перспективным использовать полученные результаты в качестве создания систем поддержки принятия решения, которые используют медицинские параметры для классификации стадий заболеваний.

Список литературы

1. Серобабов А.С. Выбор ключевых параметров для диагностики заболевания печени на основе метода анализа иерархий // Вестник кибернетики. – 2022. – №3. – С. 57-65.
2. Tukey J.W. Exploratory Data Analysis. – Reading, MA: Addison-Wesley, 1977.
3. Lee J.H., Joo L., Kang T.W. et al. Deep learning with ultrasonography: Automated classification of liver fibrosis using a deep convolutional neural network // Eur. Radiol. 2020, vol. 30, pp. 1264-1273.
4. Чубаненко Е.А., Беляева О.Д., Беркович О.А., Баранова Е.И. Значение лептина в формировании метаболического синдрома // Проблемы женского здоровья. – 2010. – №1. – С. 45-60.
5. Considine R.V., Considine E.L., Williams C.J. et al. The hypothalamic leptin receptor in humans: identification of incidental sequence polymorphisms and absence of the db/db mouse and fa/fa rat mutations // Diabetes. 1996, vol. 45, pp. 992-994.

Сведения об авторах:
Серобабов Александр Сергеевич – аспирант;
Серобабова Анастасия Леонидовна – к.т.н., младший лейтенант, инженер.