УДК 004:62-52:004.03

МЕТОДИКА НОРМАЛИЗАЦИИ НЕСВЯЗАННЫХ ДАННЫХ В УСЛОВИЯХ НЕОДНОРОДНОСТИ ВЫБОРКИ

К. т. н. Н. О. Комлевая, Д. Д. Бондаренко, М. А. Крупенко, А. С. Боренко

Одесский национальный политехнический университет
Украина, г. Одесса
nokoml@yandex.ua

Рассматриваются проблемы нормализации несвязанных данных, характеризующихся значительным уровнем рассеяния. Предлагается методика, позволяющая преобразовать эмпирические данные в совокупность данных, распределенных по нормальному закону. Исключение сильно рассеянных данных позволяет сделать выборки однородными и снизить коэффициент вариации.

Ключевые слова: неоднородность, нормальный закон распределения.

Применение простейших методов статистического анализа в производственных условиях или при экспериментальных исследованиях представляет собой выполнение четкой последовательности операций, в результате которых на основании значений статистик критериев принимается решение об отклонении или принятии проверяемых гипотез. Нормальность наблюдаемых данных является необходимой предпосылкой для корректного применения большинства классических методов математической статистики, используемых в задачах метрологии, стандартизации, классификации и контроля качества. Поэтому проверка на нормальность является обязательной процедурой в ходе проведения измерений, контроля и испытаний.

Целью данной работы является разработка методики нормализации несвязанных статистических неоднородных данных [1]. Под неоднородностью данных будем понимать некоторый уровень их рассеяния, при котором рассчитываемые статистические показатели дают надежную и качественную характеристику анализируемой совокупности. Основным мерилом разброса (и однородности) данных являются показатели вариации: дисперсия, стандартное отклонение, среднее линейное отклонение. Однако все они связаны с масштабом исходных данных и не дают «независимой» (относительной) характеристики меры разброса. Для преодоления этой проблемы используется коэффициент вариации, который рассчитывается как соотношение стандартного отклонения и средней величины. Показатель вариации не имеет единиц измерения, то есть не связан с масштабом анализируемых данных. Исходя из этого факта, коэффициенты вариации можно сравнивать между собой и тем самым сопоставлять относительную меру рассеяния данных независимо от их масштаба. В статистике принято считать, что если значение коэффициента вариации менее 33%, то совокупность данных является однородной, если более 33%, то — неоднородной.

Проблема формализации методики нормализации несвязанных статистических данных возникла в связи с тем, что большинство стандартных статистических методов анализа экспериментальных данных (дисперсионный анализ, факторный анализ, метод наименьших квадратов и т.п.) дают гарантировано правильные ответы при условии, что обрабатываемый экспериментальный материал представляет собой выборку из генеральной совокупности, распределенной по нормальному закону. В то же время конкретные статистические данные, особенно в областях медико-биологических исследований, не позволяют с достаточно большой степенью уверенности принять гипотезу о нормальности их распределения. В результате применение ряда статистических процедур зачастую становится лишь эвристическим приемом, не имеющим под собой достаточной теоретической базы.

В настоящей работе предлагается в качестве начального этапа обработки статистического материала применять процедуру нормализации, представляющую собой достаточно простое преобразование исходных данных, априорно не являющихся нормальными, в совокупность данных, распределенных по нормальному или близкому к нормальному закону.

В качестве материала для исследований была рассмотрена диагностическая группа [2, 3], которая описывалась набором векторов, каждый из которых хранил количественные значения 32 показателей, измеренных на непрерывной шкале. Диапазон значений коэффициентов вариации первоначальных эмпирических данных, вычисленных для каждой выборки, составил от 31 до 42%, что говорит о неоднородности данных.

Реально ни в одной выборке не может быть строго нормального распределения признака. Однако необходимо установить, отобрана ли эта выборка из генеральной совокупности, в которой изучаемый признак имеет гауссово распределение. Так как любая выборка, входящая в исследуемую группу, содержит меньше 60 значений, для принятия решения о виде распределения был использован достаточно мощный и универсальный критерий Шапиро—Уилка (W). Данный критерий основан на регрессии порядковых статистик на их ожидаемые значения, он позволяет снизить вероятности ошибок второго рода. С учетом критерия Шапиро—Уилка нулевая гипотеза H_0 о нормальности распределения принималась при условии, что уровень статистической значимости $p>0,05$ и W принимает высокие значения ($W>0,9$). Иначе принималась альтернативная гипотеза H_1. При проведении исследований в медико-биологических областях для повышения степени уверенности в нормальности распределения данных рекомендуется предъявлять более строгие требования к значениям p и W.

Из описывающих диагностическую группу векторов была сформирована подгруппа из 30 векторов, которые обеспечивали достаточно высокий уровень статистической значимости, при этом все значения p превышали 0,45, а значения W — 0,94. Значения показателей векторов, не включенных в подгруппу, в большей степени отличались от общего уровня. Для проверки необходимости исключения таких векторов из дальнейшей обработки требовалось применение надлежащим образом обоснованных критериев.

Для отбрасывания выделяющихся значений по данным малых выборок был использован критерием Ф. Граббса [4], основанным на отношении двух сумм квадратов отклонений. На основании соотношения расчетной величины отношения с табличной величиной при уровне значимости 5% принимается решение о том, следует ли отбросить исследуемый вектор, или же отклонение его от общего уровня может быть объяснено случайными причинами и вектор следует включить в подгруппу. С использованием данного подхода в подгруппу были включены еще 14 векторов, при этом были получены значения $p>0,32$, $W>0,92$. Для данных, взятых из подгруппы, были вычислены коэффициенты вариации. Полученный диапазон значений коэффициентов вариации составил 17 — 23%, что соответствует однородным данным.

Таким образом, приведенная методика позволила обработать данные, описывающие выбранную диагностическую группу. В процессе обработки был значительно уменьшен разброс значений показателей выборок, что позволило рассматривать эти значения как однородные данные. Доказанная нормальность полученных данных позволит в дальнейшем применять к ним большинство классических методов математической статистики.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Петри А., Сэбин К. Наглядная медицинская статистика.— М.: ГЭОТАР-Медиа, 2009.

2. Комлевая Н.О., Комлевой А.Н. Разработка информационной модели диагностирования состояния дыхательной системы // Холодильна техніка і технологія.— 2011.— Вып. 2(130).— С. 75 — 79.

3. Комлевая Н.О., Комлевой А.Н. Автоматизация диагностирования состояния дыхательной системы // Труды тринадцатой МНПК «СИЭТ-2012». — Украина, г. Одесса. — 2012. — С. 55.

4. Рябушкин Т.В., Ефимова М.Р., Ипатова И.М., Яковлева Н.И. Общая теория статистики.— М.: Финансы и статистика, 1999.

N.O. Komlevaya, D.D. Bondarenko, M.A. Krupenko, A.S. Borenko

Method of unrelated data normalization in conditions of the heterogeneous sample.

The problems of normalization of unrelated data, characterized by a significant level of scattering, are considered. The technique is proposed, which makes it possible to convert the empirical data into a set of data distributed by the normal law. The exclusion of strongly scattered data allows making samples homogeneous and reducing the coefficient of variation.

Keywords: *heterogeneity, normal distribution law.*