Для цитирования: Аналитика и контроль. 2016. Т. 20, № 1. С. 47-52
УДК 663.253.2 DOI: 10.15826/analitika.2015.20.1.009

Использование модели логит регрессии для идентификации фальсификатов вин

А.А. Халафян[1], Ю.Ф. Якуба[2], *З.А. Темердашев[1]

[1]Кубанский государственный университет, Российская Федерация, 350040, Краснодар, ул. Ставропольская, 149
[2]Северо-Кавказский зональный научно-исследовательский институт садоводства и виноградарства, Российская Федерация, 350901, Краснодар, ул. 40-лет Победы, 39

**Адрес для переписки: Темердашев Зауаль Ахлоович, E-mail: temza@kubsu.ru*

Поступила в редакцию 25 февраля 2016 г., после исправлений – 14 марта 2016 г.

Работа посвящена актуальной проблеме – оценке возможности использования метода бинарных откликов – модели логит регрессии для идентификации фальсификатов вин. При описании статистических характеристик использованной выборки образцов натуральных вин и фальсификатов наряду со средним арифметическим значением и стандартным отклонением рассмотрены их непараметрические аналоги – *медиана*, *нижняя* и *верхняя квартили*. Результаты исследований межгрупповой неоднородности натуральных вин и фальсификатов по содержанию в них летучих веществ (ацетальдегида, этилацетата, метанола, высших спиртов, уксусной кислоты, фурфурола) явились предпосылкой разработки модели логит регрессии для идентификации фальсификатов по данным химического анализа. Об адекватности построенной модели свидетельствуют результаты сравнения классификации категорий образцов, предсказанных по модели бинарных откликов с исходными категориями в выборке. Разработан алгоритм и построена программа, позволяющая по содержаниям ацетальдегида, этилацетата, метанола, суммарного содержания высших спиртов, уксусной кислоты, фурфурола автоматизировать процедуру идентификации вина (фальсификатов).

***Ключевые слова*:** натуральные вина и фальсификаты, модели логит регрессии, идентификация, описательные статистики

For citation: *Analitikaikontrol'* [Analytics and Control], 2016, vol. 20, no. 1, pp. 47-52
DOI: 10.15826/analitika.2015.20.1.009

The application of a logit regression model to identify the falsification of wines

A.A. Khalaphyan[1], Yu.F. Yakuba[2], *Z.A. Temerdashev[1]

[1]Kuban State University, 149, ul. Stavropol'skaia, Krasnodar, 350040, Russian Federation
[2]North Caucasian Regional Research Institute of Horticulture and Viticulture, 39, ul.40 let Pobedy, Krasnodar, 350901, Russian Federation

**Corresponding author: Zaual' A. Temerdashev, E-mail: temza@kubsu.ru*

Submitted 25 February 2016, received in revised form 14 March 2016

Current investigation is devoted to an actual problem – an assessment of the possibility of employing a binary responses method – logit regression model to identify the falsification of wines. The description of the statistical characteristics of the natural wines samples and counterfeits with their respective arithmetic mean values, standard deviations and nonparametric analogues – medians, inferior and top quartiles are considered. Research results for the intergroup heterogeneity of natural wines and counterfeits based on their volatiles content (acetic aldehyde, ethyl acetate, methanol, higher alcohols, acetic acid, furfural) were the prerequisite for the development of the logit regressions model to identify the falsifications by chemical analysis. The adequacy of the constructed model is demonstrated by comparing the result classification categories of the samples predicted by the model of binary responses with the original categories in the sample. The algorithm and the corresponding program are developed that allow for the values of acetic aldehyde, ethyl acetate, methanol, total contents of higher alcohols, acetic acid and furfural to be used in automating wines / counterfeits identification procedure.

Keywords: natural wines and falsificat, models a regression logit, identification, a descriptive statistics

Введение

Физико-химические показатели вин в России регламентируются национальными стандартами [1-2], устанавливающими содержание спирта, сахара, диоксида серы, титруемых и летучих кислот, токсичных элементов и радионуклидов, приведенного экстракта, лимонной кислоты. Регламентируемые этими ГОСТ испытания направлены, в основном, на контроль безопасности и позволяют установить соответствие продукции своей товарной группе, но не в полной мере дают представления о ее подлинности. Проблема качества и натуральности винодельческой продукции вызвана, в первую очередь, двумя факторами: внедрением ускоренных технологий, различных пищевых добавок и наполнителей (красителей, ароматизаторов, стабилизаторов цвета и вкуса) и практически неконтролируемым составом применяемых вспомогательных материалов. С учетом участившихся случаев отравления людей суррогатным алкоголем, особую актуальность приобретает разработка способов выявления фальсификатов.

Фальсифицированные вина обычно представляют собой искусственную смесь этилового спирта, сахарозы, лимонной, реже винной, кислоты, различных экстрактов синтетического и растительного происхождения, прочих ингредиентов и полностью соответствуют требованиям действующих государственных стандартов и СанПиН 2.3.2.1078-01 по физико-химическим показателям и критериям безопасности [3]. Однако такие «напитки», обладая плохой вкусовой характеристикой, могут стать и причиной отравлений из-за наличия нерегламентированных стандартами химических соединений, обладающих собственной токсичностью или же компонентами, усиливающими токсическое действие этилового спирта. Кроме того, существуют способы фальсификации, приводящие к улучшению органолептических свойств вина. В странах Евросоюза действует нормативно-техническая и информационно-документальная база, направленная на борьбу с некачественной и фальсифицированной продукцией [4].

В предыдущих работах [5-7] мы рассматривали возможность прогнозирования качества вин методами математической статистики при известных концентрациях определенного набора летучих и нелетучих веществ. В [5] были представлены регрессионные модели, описывающие характер стохастической взаимосвязи между содержанием аминокислот (пролина, треонина, аргинина) и дегустационной оценкой. В [6] рассмотрены взаимосвязи между дегустационной оценкой и содержаниями в виноградных винах летучих веществ, построено адекватное уравнение регрессии, при помощи которого показана возможность предсказания дегустационной оценки для вин высокого, среднего и низкого качества при известных концентрациях таких веществ, как ацетальдегид, этилацетат, метанол, высшие спирты, уксусная кислота, фурфурол. В [7] рассмотрена номинальная шкала классификации вин по категориям качества *высокое*, *среднее*, *низкое*, *фальсификат*. Показано, что классификация вин в номинальной шкале дискриминантным анализом по концентрациям указанных выше летучих веществ, обусловливающих их органолептические свойства, не уступает их экспертной (дегустационной) оценке. Построена математическая модель классификации вин по указанным категориям, разработан программный модуль для автоматизации вычислений.

В настоящей работе нами рассматривается оценка возможности использования статистического метода бинарных откликов – модели логит регрессии (логистической регрессии – logit model) для идентификации фальсификатов вин.

Материал и методы исследования

Для изучения возможности использования модели логит регрессии для идентификации фальсификатов вин по анализу состава летучих компонент использовали экспериментальные данные по химическому анализу вин, опубликованные в наших работах [6-7]. Выборка испытуемых материалов, как указывалось, была представлена 300 натуральными красными и белыми образцами виноградных вин, произведенными предприятиями Краснодарского края в 2010–2013 гг. (агрофирма «Мысхако», ОАО АПФ «Фанагория», ООО «Кубань-вино», компания «ЮВК», ООО «Вилла Виктория», винзавод «Шато Тамань», ОАО «Аврора»), а также 30 изготовленными в экспериментальных условиях купажными и искусственно сфальсифицированными винами.

Вероятностно-статистическая модель классификации вин построена в среде пакета *STATISTICA,* использован модуль «Логит регрессии» [8]. Было установлено [6], что статистически значимыми по обеим группам по описательным статистикам являются значения величин, полученные с точностью до третьего знака после запятой.

Результаты и обсуждение

Анализ эмпирических законов распределения концентрации летучих веществ в образцах натуральных вин и фальсификатов показал их несоответствие нормальному закону [7], поэтому для описания статистических характеристик выборки наряду со средним арифметическим значением (*среднее*) и стандартным отклонением (*станд. отклон.*) были рассмотрены их непараметрические аналоги – *медиана*, *нижняя* и *верхняя квартили*. Медиана соответствует такой величине переменной, левее и правее которой находится половина (50 %) ее значений. По аналогии, левее нижней или верхней квартили находится соответственно 25 % и 75 % значений переменной. Разность между верхней и нижней квартилью – квартильный размах, явля-

Таблица 1

Описательные статистики содержания летучих веществ в натуральных винах, мг/дм^3

Descriptive statistics of contents of volatile substances in natural wines (mg/dm^3)

Переменная	Описательные статистики						
	Среднее	Медиана	Минимум	Максимум	Нижняя квартиль	Верхняя квартиль	Станд. отклон.
Ацетальдегид	75.513	44.000	21.000	230.000	35.000	126.000	58.577
Этилацетат	71.260	58.000	44.000	175.000	54.000	68.500	31.264
Метанол	84.740	86.500	40.000	165.000	54.000	120.000	33.645
Высшие спирты	318.363	258.000	200.000	750.000	241.000	340.500	122.838
Уксусная кислота	498.393	408.000	240.000	900.000	350.000	598.000	172.465
Фурфурол	30.780	20.000	2.000	97.000	6.000	56.000	26.871

Таблица 2

Описательные статистики содержания летучих веществ в фальсификатах, мг/дм^3

Descriptive statistics of contents of volatile substances in falsificat wines (mg/dm^3)

Переменная	Описательные статистики						
	Среднее	Медиана	Минимум	Максимум	Нижняя квартиль	Верхняя Квартиль	Станд. отклон.
Ацетальдегид	6.533	6.000	1.000	12.000	4.000	9.000	3.104
Этилацетат	15.533	14.000	5.000	27.000	11.000	22.000	6.709
Метанол	18.700	11.000	3.000	82.000	8.000	22.000	18.430
Высшие спирты	42.000	39.000	9.000	90.000	22.000	62.000	22.230
Уксусная кислота	91.633	90.000	5.000	210.000	65.000	112.000	48.331
Фурфурол	2.600	2.000	0.000	8.000	1.000	4.000	2.343

ется мерой разброса величин, имеющих асимметричное распределение.

Рассчитанные по содержаниям летучих компонент в натуральных виноградных винах и фальсификатах значения основных описательных статистик (экспериментальные данные взяты из [6, 7]) отображены в табл. 1 и 2. Значительное отличие медиан и средних у исследуемых образцов показывает наличие асимметричности эмпирических распределений, что является дополнительным подтверждением их несоответствия нормальному закону. Как видно, статистические характеристики содержаний компонентов существенно отличаются у натуральных вин и фальсификатов.

Для визуального представления степени отличия медиан, нижних и верхних квартилей, а также минимальных и максимальных значений содержаний компонентов у натуральных вин и фальсификатов в качестве примера на рис. 1 и 2 представлены диаграммы размаха массовых концентраций ацетальдегида и этилацетата. Из диаграмм видно, что медианы концентраций летучих веществ натуральных вин расположены значительно выше, чем медианы концентраций летучих веществ фальси-

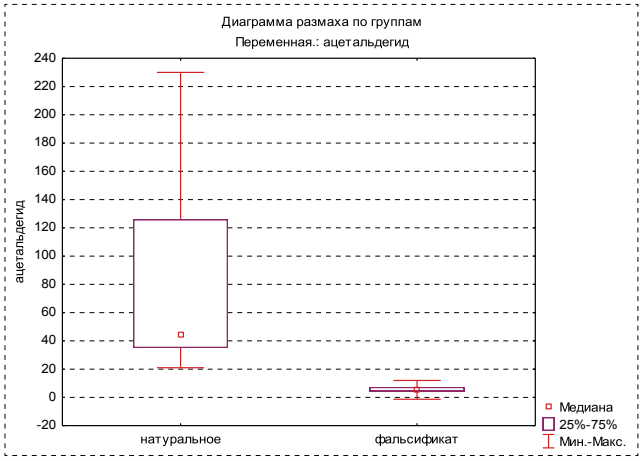

Рис. 1. Диаграмма размаха содержаний ацетальдегида в натуральных винах и фальсификатах

Fig. 1. Diagram of range of the maintenances of acetic aldehyde in natural wines and falsificats

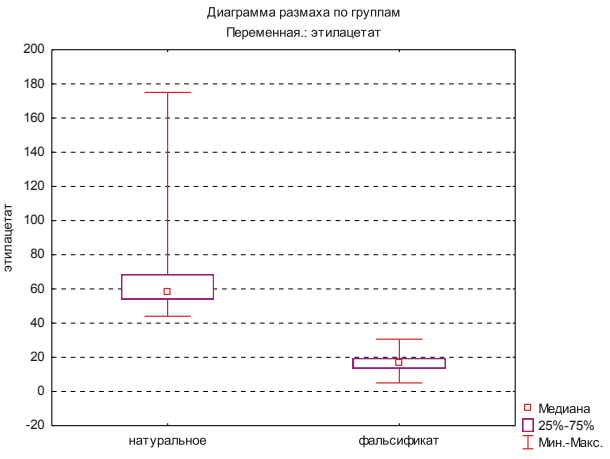

Рис. 2. Диаграмма размаха содержаний этилацетата в натуральных винах и фальсификатах

Fig. 2. Diagram of range of the maintenance of ethyl acetate in natural wines and falsificats

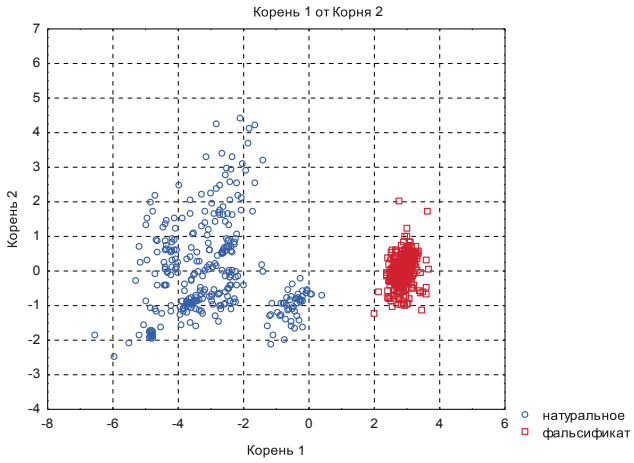

Рис. 3. Диаграмма рассеяния канонических корней для натуральных вин

и фальсификатов

Fig. 3. Diagram of dispersion of the canonical roots for natural wines and falsificats

фикатов. Квартальные размахи в значениях концентраций для натуральных вин существенно выше, чем разбросы в значениях концентраций для фальсификатов.

Столь значительные отличия (в несколько раз) средних и медиан концентраций летучих веществ натуральных вин и фальсификатов не требуют дополнительных исследований по оценке их статистической значимости. Так, например, у ацетальдегида и этилацетата средние и медианы натуральных вин и фальсификатов отличаются более чем в 4 раза, а у фурфурола – более чем в 10 раз. Диаграмма рассеяния канонических корней на рис. 3, построенная модулем *Дискриминантный анализ* программы *STATISTICA* путем визуализации, характеризует степень межгрупповых отличий по совокупности содержаний летучих веществ. На диаграмме образцы вин для каждой группы изображены в виде одинаковых геометрических фигурок одного цвета: кружочек (синий) соответствует натуральным винам, квадратик (красный) – фальсификатам.

Диаграмма рассеяния канонических корней позволяет образцы вин, являющиеся объектами в шестимерном пространстве (по количеству летучих веществ), перенести в пространство размерности 2, сохранив порядок расстояний между ними. Чем меньше расстояние между геометрическими фигурками, изображающими образцы вин, тем больше сходство между ними по совокупности веществ, и наоборот, чем больше расстояние, тем более они различны. Из диаграммы видно, что кластеры, соответствующие натуральным винам и фальсификатам локализованы в разных частях плоскости на значительном расстоянии друг от друга, что говорит об их существенном межгрупповом отличии и высоком сходстве (однородности) образцов вин внутри каждой группы. Некоторая неоднородность присутствует у натуральных вин – группа вин высокого качества образует самостоятельный кластер, расположенный между значениями – 2 и 0 канонического корня 1.

Представленные выше результаты служат предпосылкой к использованию метода бинарных откликов для идентификации фальсификатов по данным химического анализа [6, 7].

Бинарные модели применяют, когда зависимая переменная (отклик) может принимать только два значения, т.е., в тех случаях, когда представляет интерес поиск зависимостей между одной или несколькими непрерывными переменными (предикторами) и одной зависимой от них бинарной переменной со значениями 0 и 1 (в нашем случае – натуральное вино и фальсификат). Технически достаточно сложно смоделировать бинарную функцию от непрерывных предикторов, поэтому задачу регрессии формулируют иначе. Предсказывают непрерывную переменную Y из отрезка [0, 1]. При этом, если Y меньше, чем 0.5, то бинарной переменной присваивают значение 0, в противном случае – присваивают 1.

В модели логит регрессии для вычисления Y применяют логистическую функцию:

$$Y = \frac{e^Z}{1 + e^Z}, \qquad (1)$$

где $Z = b_0 + b_1X_1 + ... + b_nX_n$. Коэффициенты при $b_i (i = 1, ..., n)$ находят по обучающей выборке, для которой известна принадлежность исследуемых объектов к группам, задаваемым бинарной переменной. В нашем случае – это 330 образцов с известной принадлежностью к натуральным винам, или фальсификатам. Логистическая функция обладает тем замечательным свойством, что вне зависимости от коэффициентов регрессии и величин X_i значения отклика Y всегда будут принадлежать отрезку [0, 1].

Для вычисления коэффициентов b_0 b_1, ... b_n использовали процедуру *логит регрессии* модуля *Нелинейное оценивание* программы *STATISTICA*. Метод Хука-Дживиса в комбинации с квази-ньютоновским при начальных значениях коэффициентов, равных 0,0 и начальном размере шага 1, позволил

Таблица 3

Итоговые значения процедуры логит регрессия (итоговые потери: 0.000000195; Xu^2 (6) = 831.78; p = 0.000; n = 330)

Total value of procedure of regression logit (total loss: 0.000000195; Xu^2 (6) = 831.78; p = 0.000; n = 330)

Коэффициенты	b_0	b_1	b_2	b_3	b_4	b_5	b_6
	47.872	– 1.277	– 0.255	– 0.015	– 0.045	– 0.035	– 0.717

построить адекватное уравнение регрессии. Итоги построения модулем модели логит регрессии отображены в табл. 3.

Об адекватности построенной модели свидетельствуют близкие к 0 итоговые потери (0.000000195), оцененные функцией максимального правдоподобия и уровень значимости *p* критерия *Хи-квадрат*, который меньше 0.05. Коэффициенты b_i (i = 1, …, 6) при концентрациях летучих веществ составили соответственно: для ацетальдегида b_1 = –1.277; этилацетата b_2 = – 0.255; метанола b_3 = – 0.015; высших спиртов b_4 = – 0.045; уксусной кислоты b_5 = – 0.035 и фурфурола b_6 = – 0.717; свободный член уравнения b_0 = 47.872. Здесь следует пояснить, что коэффициенты для конкретных веществ с номером имеют отношение к записи математической модели, представленной уравнением (1), а концентрации для этих же веществ – к реальным обозначениям предикторов X_i модели в виде концентраций *C*.

С учетом значений коэффициентов b_i уравнение для вычисления *Z* примет вид:

$$Z = 47.872 - 1.277C_а - 0.255C_э - 0.015C_м + 0.045C_{вс} - 0.035C_у - 0.717C_ф, \quad (2)$$

где $C_а$, $C_э$, $C_м$, $C_{вс}$, $C_у$ и $C_ф$ – массовые концентрации ацетальдегида, этилацетата, метанола, суммарного содержания высших спиртов, уксусной кислоты и фурфурола, в мг/дм^3.

Чтобы по значениям предикторов сделать прогноз бинарного отклика, надо подставить в выражение (2) числовые значения переменных $C_а$, $C_э$, $C_м$, $C_{вс}$, $C_у$ и $C_ф$, вычислить *Z*, далее по уравнению (1) вычислить *Y*. Округлив полученное значение до целого, естественно, это будет либо 0 – натуральное вино, либо 1 – фальсификат, получим прогнозное значение отклика.

Об адекватности построенной модели также свидетельствуют результаты сравнения классификации категорий образцов, предсказанных по модели бинарных откликов с исходными категориями образцов в выборке – в обеих группах на 100 % достигнута верная классификация.

Для иллюстрации процедуры вычислений по логит регрессионной модели воспользуемся экспериментальными данными для двух образцов из рассмотренной нами выборки.

Пример 1. Определим категорию образца при следующих значениях концентраций летучих веществ ($C_а$ = 80, $C_э$ = 110, $C_м$ = 95, $C_{вс}$ = 500, $C_у$ = 650, $C_ф$ = 50 мг/дм^3):

Z = 47.872 – 1.277·80 – 0.255·110 – 0.015·95 + 0.045·500–0.035·650–0.717·50 = – 119.863.

В этом случае $Y = e^{-154.312}/(1+ e^{-154.312}) \approx 0$, т.е. образец является натуральным вином.

Пример 2. Определим категорию образца при следующих значениях концентраций летучих веществ ($C_а$ = 8, $C_э$ = 20, $C_м$ = 35, $C_{вс}$ = 60, $C_у$ = 110, $C_ф$ = 5 мг/дм^3):

Z = 47.872 – 1.277·8 – 0.255·20 – 0.015·35 + 0.045·60–0.035·110–0.717·5 = 27.296.

В этом случае $Y = e^{21,347}/(1+ e^{21,347}) \approx 1$, т.е. образец является фальсификатом.

Логит регрессия обладает недостатком, присущим всем регрессионным моделям. Она предсказывает достоверные значения отклика – бинарной переменной, если значения предикторов принадлежат диапазонам их изменения, по которым составлена модель. Поэтому, с учетом данных табл. 1 и 2, задачу идентификации фальсификатов целесообразно решать, если $C_а \in [1, 230]$, $C_э \in [5, 175]$, $C_м \in [3, 165]$, $C_{вс} \in [9, 750]$, $C_у \in [5, 900]$, $C_ф \in [0, 97]$ (мг/дм^3).

С целью автоматизации вычислительной процедуры в соответствии с уравнениями (1) и (2) была написана программа на языке Delphi, диалоговое окно которой представлено на рис. 4. Для идентификации фальсификата (натурального вина) в поля окна достаточно ввести значения массовых концентраций летучих веществ (введены данные из примера 2) и щелкнуть по кнопке *Определить категорию образца*. В нижней части окна появится сообщение *фальсификат*, или *натуральное вино*.

Таким образом, на примере описанной выборки показана возможность идентификации натуральных вин и фальсификатов при помощи модели бинарных откликов – логит регрессии. По выборке, состоящей из 330 образцов, построена математическая модель, которая позволяет по концентрации ацетальдегида, этилацетата, метанола, суммарного содержания высших спиртов, уксусной кислоты, фурфурола определить натуральность вин. По аналогии, выявив вещества, наиболее полно описывающие принадлежность образцов к определенным видам алкогольной продукции, можно при помощи модели логит регрессии решить задачу выявления их фальсификатов.

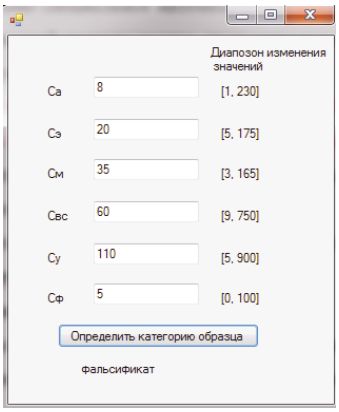

Рис. 4. Окно программы идентификации фальсификата (натурального вина)

Fig. 4. Window of program for identification of falsificat (natural wine)

ЛИТЕРАТУРА

1. ГОСТ Р 32030-2013 «Вина столовые и виноматериалы столовые. Общие технические условия». 12 с.
2. ГОСТ Р 55242-2012 «Вина защищенных географических указаний и вина защищенных наименований места происхождения. Общие технические условия». 12 с.
3. СанПиН 2.3.2.1078-01. Гигиенические требования безопасности и пищевой ценности пищевых продуктов. Москва. 2011. 145 с.
4. [Электронный ресурс]: http://www.oiv.int/oiv/info/en-methodesinternationalesvin/International Methods of Analysis of Wines and Musts (дата обращения 14.07.2015).
5. Якуба Ю.Ф., Темердашев З.А., Халафян А.А. Вкусовая оценка качества виноградных вин с использованием методов математической статистики // Вопросы питания. 2016. Т.85, № 2. С. 17-25.
6. Якуба Ю.Ф., Темердашев З.А., Халафян А.А. Органолептическая оценка качества виноградных вин с использованием методов статистического моделирования // Аналитика и контроль. 2014. Т. 18, № 4. С. 385-392.
7. Якуба Ю.Ф., Халафян А.А., Темердашев З.А. Применение классификационного анализа для оценки качества вин в номинальной шкале // Журнал аналитической химии. 2016. Т.71, № 2. С. 212-222.
8. Халафян А.А. STATISTICA 6. Математическая статистика с элементами теории вероятностей. М.: Бином, 2010. 491 с.

REFERENCES

1.*GOST R 32030-2013. Vina stolovye i vinomaterialy stolovye. Obshchie tekhnicheskie usloviia* [State Standard 32030-2013. Table wines and wine materials. General specifications]. Moscow, Standartinform Publ., 2013. 12 p. (in Russian).
2. GOST R 55242-2012. *Vina zashchishchennykh geograficheskikh ukazanii i vina zashchishchennykh naimenovanii mesta proiskhozhdeniia. Obshchie tekhnicheskie usloviia* [State Standard 55242-2012.Wines from protected geographical indications and wines with a protected place of origin. General specifications]. Moscow, Standartinform Publ., 2013. 12 p. (in Russian).
3. SanPiN 2.3.2.1078–01. *Gigienicheskie trebovaniia bezopasnosti i pishchevoi tsennosti pishchevykh produktov* [Sanitary Standard 2.3.2.1078–01. Hygienic safety and nutritional value of foods]. Moscow. 2011. 144 p. (in Russian).
4. *International Methods of Analysis of Wines and Musts (2015*). Available at: http://www.oiv.int/oiv/info/enmethodesinternationalesvin OIV (accessed 14 July 2015).
5.Yakuba Yu.F., Temerdashev Z.A., Khalaphyan A.A. [Testing estimation of quality of grape wines with use of methods of statistic mathematical]. *Voprosy pitaniia* [Questions of nutrition], 2016, vol. 85, no. 2, pp. 17-25. (in Russian).
6. Yakuba Yu.F., Temerdashev Z.A., Khalaphyan A.A. [Organoleptic estimation of quality of grape wines with use of methods of statistical modeling]. *Analytics and control*, 2014, vol. 18, no. 4, pp. 385-392. (in Russian).
7. Yakuba Yu.F., Temerdashev Z.A., Khalaphyan A.A. Application of ranging analysis to the quality assessment of wines on a nominal scale. Journal of analytical chemistry, 2016, vol.71, no. 2, pp. 205-214. DOI: 10.1134/S1061934816020155
8. Khalaphyan A.A. *STATISTICA 6. Matematicheskaia statistika s elementami teorii veroiatnostei* [STATISTICA 6. Mathematical statistics with elements of theory of probability]. Moscow, Binom, 2010, 491 p.