ISSN 2713-1734

Научная статья
УДК 519.254
EDN AGZLFU
DOI 10.17150/2713-1734.2023.5(4).476-493

Н.М. Борбаць
Брянский государственный технический университет,
г. Брянск, Российская Федерация
Т.В. Школина
Брянский государственный технический университет,
г. Брянск, Российская Федерация

Процедура подбора кривой из системы Джонсона методами процентилей и максимального правдоподобия — наименьших квадратов в R

Аннотация. Описывается реализация и применение процедуры на языке R, предназначенной для подбора кривой распределения из системы Джонсона. В предлагаемой процедуре, в отличие от уже реализованной функции из пакета «SuppDists», выбор семейства кривых и оценка параметров осуществляется с использованием метода процентилей и комбинации методов максимального правдоподобия и наименьших квадратов. Возможность устанавливать начальные значения вектора квантилей стандартного нормального распределения позволяет пользователю осуществлять подбор кривой более гибко. Кроме того, оценка качества подбора может быть выполнена визуально, с использованием графиков функций плотности и функций распределения, а также формализовано с применение критерия согласия хи-квадрат Пирсона.

Ключевые слова. Кривые Джонсона, подбор распределения, язык R.

Информация о статье. Дата поступления: 4 апреля 2023 г.; дата принятия к публикации: 20 ноября 2023 г.; дата онлайн-размещения: 12 декабря 2023 г.

Original article

N.M. Borbats
Bryansk State Technical University,
Bryansk, Russian Federation
T.V. Shkolina
Bryansk State Technical University,
Bryansk, Russian Federation

The Procedure for Selecting a Curve from the Johnson System by Percentile Matching and Maximum Likelihood and Least Squares Approaches in R

Abstract. It describes the implementation and application of a procedure in the R language designed to select a distribution curve from the Johnson system. In the proposed procedure, unlike the already implemented function from the SuppDists package, the selection of a family of curves and the estimation of parameters is carried out using the percentile method and a combination of maximum likelihood and least squares methods. The ability to set the initial values of the quantile vector of the standard normal distribution allows the user to select the curve more flexibly. In addition, the assessment of the quality of the selection can be performed visually, using graphs of density func-

tions and distribution functions, as well as formalized with the application of Pearson's chi-squared agreement criterion.

Keywords. Johnson curves, selection of distribution, language R.

Article info. Received 20 June, 2023; Accepted 20 November, 2023; Available online 12 December, 2023.

Введение

При применении большинства статистических инструментов, используемых для принятия решений в технических системах, предполагается, что распределение контролируемой непрерывной случайной величины является нормальным. Однако зачастую фактическое распределение может сильно отличатся от нормального, и для применения соответствующего статистического метода встает задача выбора вероятностного распределения, в наибольшей степени соответствующего эмпирическому распределению контролируемой характеристики X.

При этом, выбор модели распределения значений рассматриваемой характеристики должен основываться на понимании физической природы соответствующего процесса или системы и не должен осуществляться простым перебором большого числа различных статистических моделей и принятии той из них, которая обеспечивает наилучшее соответствие результатам наблюдений по результатам применения используемого критерия согласия [1]. Однако, зачастую не имеется достаточно надежных теоретических оснований для выбора какой-либо статистической модели распределения вероятностей. В таком случае выходом может быть аппроксимация данных эмпирическим распределением с использованием одной из двух наиболее часто применяемых на практике систем кривых — системы кривых Джонсона или системы кривых Пирсона [2].

Наиболее простой из двух указанных систем является система кривых Джонсона, основанная на преобразовании стандартной нормальной случайной величины и охватывающая широкий диапазон распределений различной формы [3]. Подбор эмпирического распределения в системе Джонсона на практике предполагает решение двух основных задач: 1) выбор семейства распределений, и 2) оценка параметров кривой выбранного семейства. Очевидно, решение обеих задач целесообразно выполнять с применением соответствующего программного обеспечения, примером которого может служить язык программирования R, одним из достоинств которого является то, что он представляет собой свободно распространяемое программное обеспечение с открытым исходным кодом[1].

[1] R Development Core Team. R: A language and environment for statistical computing. — R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/.

Далее в разделе 1 излагаются краткие сведения о системе кривых Джонсона и особенностях входящих в нее семейств распределений. В разделе 2 дается обзор основных методов выбора семейства распределения и оценки его параметров. Раздел 3 посвящен описанию предлагаемой процедуры на языке R, предназначенной для выбора семейства распределения и оценки параметров, включая описание параметров процедуры и ее выходных данных. В разделе 4 приводятся примеры применения описываемой процедуры.

1. Система кривых Джонсона

Пусть X — непрерывная случайная величина с неизвестным распределением, для которой требуется подобрать эмпирическое распределение. Джонсоном была предложена система, включающая три семейства распределений, которые в общем виде задаются преобразованием вида [3]:

$$z = \gamma + \eta\tau(x; \varepsilon, \lambda),$$
$$\eta > 0, -\infty < \gamma < +\infty, \lambda > 0, -\infty < \varepsilon < +\infty,$$

где z — стандартная нормальная случайная величина; η, γ, λ и ε — параметры распределения; $\tau(x; \varepsilon, \lambda)$ — произвольная функция;

Джонсон предложил следующие три формы функций τ:

$$\tau_1(x; \varepsilon, \lambda) = \ln\left(\frac{x-\varepsilon}{\lambda}\right) \quad \text{при} \quad x \geq \varepsilon,$$

которая соответствует логнормальному семейству распределений и обозначается как S_L;

$$\tau_2(x; \varepsilon, \lambda) = \ln\left(\frac{x-\varepsilon}{\lambda+\varepsilon-x}\right) \quad \text{при} \quad \varepsilon \leq x \leq \varepsilon+\lambda,$$

соответствующая семейству распределений случайных величин, имеющих ограниченную область возможных значений, и обозначаемая как S_B;

$$\tau_3(x; \varepsilon, \lambda) = \sinh^{-1}\left(\frac{x-\varepsilon}{\lambda}\right) \quad \text{при} \quad -\infty \leq x \leq +\infty,$$

которая соответствует семейству распределений неограниченных случайных величин и обозначается как S_U.

Кривые семейства S_L по сути могут быть заданы только тремя параметрами, так как параметр λ может быть исключен путем задания

$$\gamma^* = \gamma - \eta\ln(\lambda),$$

так что

$$z = \gamma^* + \eta\ln(x - \varepsilon).$$

Учитывая, что в выражении (1) z является стандартной нормальной величиной, используя методы преобразования случайных величин [1], можно получить формулы для функций плотности распределения вероятностей кривых каждого из трех семейств. При этом для удобства их записи часто вводится вспомогательная переменная:

$$\gamma = (x - \varepsilon) / \lambda,$$

тогда функции τ соответствующих семейств могут быть представлены в виде:

$$g(y)=\begin{cases}\ln(y) & \text{для} \quad S_L;\\ \ln\left(\dfrac{y}{1-y}\right) & \text{для} \quad S_B;\\ \ln\left(y+\sqrt{y^2+1}\right) & \text{для} \quad S_U.\end{cases}$$

2. Основные методы выбора семейства кривых и оценки параметров

Существуют различные методы выбора семейства кривых Джонсона и последующей оценки параметров кривой выбранного семейства для аппроксимации экспериментальных данных. К числу основных можно отнести:

1. Метод моментов.
2. Метод квантилей.
3. Метод процентилей.

Выбор семейства кривой на основе моментов основан на том факте, что на графике стандартизированных моментов третьего и четвертого порядка $\sqrt{\beta_1}$ и β_2 кривая, соответствующая распределению семейства S_L, делит плоскость β_1, β_2 на две области. При этом кривые распределений семейства S_B лежат в одной области, а кривые распределений семейства S_U — в другой [3]. Тогда, вычислив по экспериментальным данным значения оценок стандартизированных моментов и определив в какую из двух областей попадает соответствующая им точка, можно выбрать семейство распределений. В свою очередь, процедура оценки параметров кривой выбранного семейства на основе метода моментов предполагает приравнивание первых четырех выборочных моментов, найденных по данным, к соответствующим теоретическим моментам выбранного семейства и последующего решения полученных уравнений [3, 4–6].

Однако практическая реализация метода моментов обладает радом недостатков, основными из которых являются [7]:

– выборочные оценки центральных моментов третьего и четвертого порядка имеют достаточно большие дисперсии;

– выборочные оценки этих моментов оказываются сильно смещенными при малых размерах выборок;

– оценки моментов сильно подвержены выбросам в данных.

В этой связи практически одновременно были предложены два других метода выбора семейства кривых и оценки их параметров — метод квантилей и метод процентилей. В соответствии с методом квантилей [8] для заданного значения кумулятивной вероятности p_n определяют соответствующее ей значение квантиля стандартного нормального распределения z_n. После этого выбирают пять эмпирических квантилей из данных x_p, x_k, x_0, x_m, x_n, соответствующих квантилям стандартного нормального распределения: $z = -z_n$, $-0{,}5z_n$, 0, $0{,}5z_n$, z_n. Тогда для выбора семейства кривых может использоваться соотношение:

$$\frac{t_b}{t_u} = \frac{(x_m - x_0)(x_m - x_k)}{(x_n - x_m)(x_0 - x_p)}.$$

Можно показать, что $t_b / t_u < 1$ для распределений из семейства S_U, $t_b / t_u > 1$ для распределений из семейства S_B и $t_b / t_u = 1$ для распределений из семейства S_L. После выбора семейства кривых распределения Джонсона с использованием найденных ранее эмпирических квантилей x_p, x_k, x_0, x_m, x_n вычисляются оценки параметров кривой.

При методе процентилей [7] выбирается какое-либо значение $z > 0$ стандартной нормальной величины. На основе выбранного значения z определяются три интервала области значений стандартной нормальной случайной величины равной длины — в самом простом случае, для этого можно использовать четыре точки: $-3z$, $-z$, $+z$ и $+3z$. Для заданных четырех точек определяются соответствующие им эмпирические процентили x_{-3z}, x_{-z}, x_z, и x_{3z}: для этого вначале необходимо найти значения кумулятивных вероятностей стандартного нормального закона P_ζ, соответствующие значениям $\zeta = -3z$, $-z$, $+z$ и $+3z$, т.е. $P_\zeta = \Phi(\zeta)$, где $\Phi(\zeta)$ — функция распределения стандартного нормального закона. Тогда эмпирическим квантилем x_ζ, соответствующим значению ζ, будет является i-я порядковая статистика, т.е. $x_\zeta = x_{(i)}$, где $i = NP_\zeta + 1/2$, а N — объем данных. При этом, поскольку получаемые значения i в общем случае будут нецелыми, то для нахождения $x_{(i)}$ необходимо использовать интерполяцию.

На основе найденных эмпирических квантилей x_{-3z}, x_{-z}, x_z, и x_{3z} выбирается семейство кривых, для чего используется отношение mn / p^2, где:

$$\begin{aligned} m &= x_{3z} - x_z; \\ n &= x_{-z} - x_{-3z}; \\ p &= x_z - x_{-z}. \end{aligned} \qquad (2)$$

Можно показать, что $mn / p^2 > 1$ соответствует любому распределению семейства S_U, $mn / p^2 < 1$ — любому распределению семейства S_B, а $mn / p^2 = 1$ — любому распределению семейства S_L. Тогда, задавшись допустимой ошибкой сравнения Δ, можно использовать следующее правило выбора семейства кривых:

– если $mn / p^2 > 1 + \Delta$, то выбирается семейство распределений S_U;

– если $mn / p^2 < 1 - \Delta$, то выбирается семейство распределений S_B;

– если $\left|mn / p^2 - 1\right| \le \Delta$, то выбирается семейство распределений S_L.

После выбора семейства кривых распределения Джонсона значения m, n и p используются для нахождения параметров соответствующей кривой.

Наконец, относительно недавно был предложен новый алгоритм оценки параметров распределений Джонсона, основанный на совместном применении методов максимального правдоподобия и наименьших квадратов [9] — в дальнейшем метод MLE-LS. Процедура оценки параметров данным методом носит итерационный характер и начинается с некоторых начальных значений $\hat{\varepsilon}$ и $\hat{\lambda}$, на основе которых, используя соответствующие зависимости, находят оценки параметров $\hat{\eta}$ и $\hat{\gamma}$ (или $\hat{\gamma}^*$ для семейства S_L). После нахождения оценок параметров $\hat{\eta}$ и $\hat{\gamma}$ (или $\hat{\gamma}^*$) пересматривают оценки $\hat{\varepsilon}$ и $\hat{\lambda}$, после чего процедура повторяется. Поскольку при заданных значениях параметров $\hat{\eta}$ и $\hat{\gamma}$ (или $\hat{\gamma}^*$) параметры ε и λ оцениваются как параметры парной линейной регрессии, то после нахождения соответствующих оценок $\hat{\varepsilon}$ и $\hat{\lambda}$ вычисляется сумма квадратов остатков (RSS), минимизация которой и является условием остановки процедуры оценки.

Поскольку метод MLE-LS основан на заданных начальных значениях оценок $\hat{\varepsilon}$ и $\hat{\lambda}$, то для их получения, а также для выбора семейства кривых, может использоваться один из других методов. На основе проведенного имитационного исследования было установлено, что оценки параметров, найденные методом MLE-LS, в среднем ближе к своим истинным значениям, а средний квадрат ошибки в общем случае меньше, чем при использовании других методов.

3. Процедура выбора семейства кривых и оценки параметров в R

На данный момент аппроксимация результатов наблюдений эмпирическим распределением из системы Джонсона в R может быть выполнена с помощью пакета «SuppDists» [10]. Данный пакет включает в себя помимо прочего функции для расчета значений плотности, функции распределения, квантилей, генерации случай-

ных чисел для распределений из семейств Джонсона с заданными параметрами, а также подбора кривой и оценки ее параметров методом квантилей и моментов.

В качестве дополнения к соответствующей функции из пакета «SuppDists» авторами разработана функция для подбора кривой системы Джонсона на основе метода процентилей и метода MLE-LS, полный листинг которой доступен по ссылке https://goo-gl.me/b076i. Общий синтаксис функции имеет вид:

```
fit_Johnson(x, msp = "percentile", method = "MLE-LS",
            z = 0.5, alpha = NULL, tol = 1e-4, iter = 100,
            p = NULL, plot = FALSE, chi2test = FALSE)
```

Аргументами функции являются:

x — числовой вектор выборочных данных, которые необходимо аппроксимировать кривой из системы Джонсона. Если в векторе содержатся пропущенные значения (NA), то в процессе дальнейшей обработки они удаляются;

method — строка символов, указывающая метод оценки параметров подбираемой кривой, по умолчанию установлено значение "MLE-LS", т.е. оценки параметров кривой из выбранного семейства находятся методом MLE-LS. Второе допустимое значение — "percentile", в этом случае возвращаются значения параметров, найденные методом процентилей;

z — исходное значение квантиля стандартного нормального распределения, используемое при оценке параметров подбираемой кривой методом процентилей, по умолчанию используется значение 0,5. На основе заданного исходного значения определяются остальные три квантиля так, что они оказываются равноудаленными друг от друга, а именно: ($-3z$, $-z$, $+z$ и $+3z$), при этом их среднее равно нулю. Если при использовании метода процентилей требуется обеспечить лучшую аппроксимацию к данным на одном из хвостов распределения, то в качестве аргумента z можно задать вектор из четырех равноудаленных значений, которые соответствуют квантилям стандартного нормального распределения, среднее которых не равно нулю;

alpha — значение порядка квантиля стандартного нормального распределения, соответствующего значению $-3z$ в методе процентилей. Например, вместо задания аргумента z, равного 0,5, можно задать значение аргумента alpha, примерно равное 0,067 — результат будет одинаковым. Следует учитывать, что так как между параметрами alpha и z существует взаимосвязь, то при задании одного параметра, второй игнорируется с предупреждением, при этом если z — скаляр, то при заданном параметре alpha значение z игнорируется, если же z — четырехэлементный вектор, то иг-

норируется параметр alpha в случае его задания. Кроме того, при использовании в методе MLE-LS начальных значений оценок параметров ε и λ, полученных методом квантилей или моментов, значения обоих аргументов z и alpha, отличные от установленных по умолчанию, игнорируются с предупреждением;

msp — метод для определения семейства кривых и начальных значений оценок параметров. Доступными значениями являются: "percentile" — для метода процентилей, "quant" — для метода квантилей и "moment" — для метода моментов. Следует учитывать, что значения данного аргумента имеют смысл, только если для аргумента method заданно значение "MLE-LS", при method = "percentile" все значения аргумента msp кроме "percentile" игнорируются с предупреждением;

p — значение или вектор значений уровней квантилей подобранной кривой из семейства Джонсона, которые необходимо вернуть;

iter — число итераций, выполняемых при оценке параметров подбираемой кривой методом MLE-LS. Если процесс оценки не завершается успехом до выполнения указанного числа итераций, возвращаются наилучшие из найденных оценок, с предупреждением;

tol — допустимая ошибка сравнения Δ при выборе семейства кривых методом процентилей, по умолчанию установлено значение 10^{-4};

plot — логический аргумент, указывающий необходимость выполнять графическое сопоставление распределения выборочных данных с подобранным распределением из системы Джонсона. Если установлено значение TRUE, то в графическое устройство выводится гистограмма с наложенным на нее графиком функции плотности подобранной кривой и графики эмпирической и теоретической кумулятивных функций (cdf);

chi2test — логический аргумент, указывающий нужно ли проводить оценку качества подбора по критерию согласия хи-квадрат Пирсона. При этом, так как применение критерия согласия хи-квадрат предполагает, по крайней мере, умеренно большой объем данных, то задание для данного аргумента значения TRUE при объеме данных $n < 50$ приведет к сообщению об ошибке и завершению работы процедуры.

Выходным результатом функции является список со следующими компонентами:

type — строка символов, указывающая на выбранное семейство кривых Джонсона;

coefficients — именованный вектор с элементами, eta, gamma, lambda и epsilon, соответствующих оценкам параметров кривой из выбранного семейства;

RSS — остаточная сумма квадратов, соответствующая итерации, на которой были найдены оценки параметров методом

MLE-LS. Если для входного аргумента method заданно значение "percentile", то данному компоненту присваивается значение NA;

method — строка символов "Percentile Matching" или "MLE-Least Square Approach", указывающая на использованный метод нахождения оценок параметров кривой;

percentiles — именованный вектор, содержащий значения процентилей подобранного распределения Джонсона для уровней, указанных в аргументе p. Если аргумент p не задан (по умолчанию), данному компоненту присваивается значение NA.

Применение описываемой функции требует предварительно установленных пакетов «ggplot2» и «SuppDists». Общий алгоритм работы функции приведен на рис. 1 и предполагает использование следующих вспомогательных процедур, оформленных в виде подфункций:

est_percentile — используется для оценки параметров кривой методом процентилей на основе выбранного семейства, найденных по формулам (2) значений m, n и p, и заданного вектора значений $-3z$, $-z$, $+z$ и $+3z$. В данной процедуре используется несколько измененный алгоритм, реализованный в MATLAB;

get_quantiles — вспомогательная функция для определения на основе входного аргумента x эмпирических квантилей x и квантилей стандартного нормального распределения z, используемых при оценке параметров подбираемой кривой методом MLE-LS;

est_MLESL — функция оценки параметров подбираемой кривой в системе Джонсона с использованием метода MLE-LS на каждой отдельной итерации. Входными аргументами функции являются: дата-фрейм, содержащий значения эмпирических квантилей и соответствующих им квантилей стандартного нормального распределения, возвращаемый процедурой get_quantiles; начальные значения оценок $\hat{\varepsilon}$ и $\hat{\lambda}$; указание на семейство подбираемой кривой;

make_chisqtest — вспомогательная функция для выполнения проверки качества подбора по критерию согласия хи-квадрат Пирсона. Процедура основана на использовании встроенной статистической функции chisq.test для сравнения наблюдаемых частот и заданных (теоретических) вероятностей. Входными аргументами процедуры являются: наблюдаемые частоты, получаемые в результате группировки выборочных данных; значения верхних границ интервалов группировки; число оценок параметров подобранной кривой. Теоретические вероятности определяются на основе найденных оценок параметров с использованием функции pJohnson из пакета «SuppDists». Поскольку применение критерия согласия хи-квадрат требует, чтобы все ожидаемые (теоретические) частоты были не меньше пяти, то перед вызовом функции chisq.test проверяется выполнение этого требования и, при необходимости,

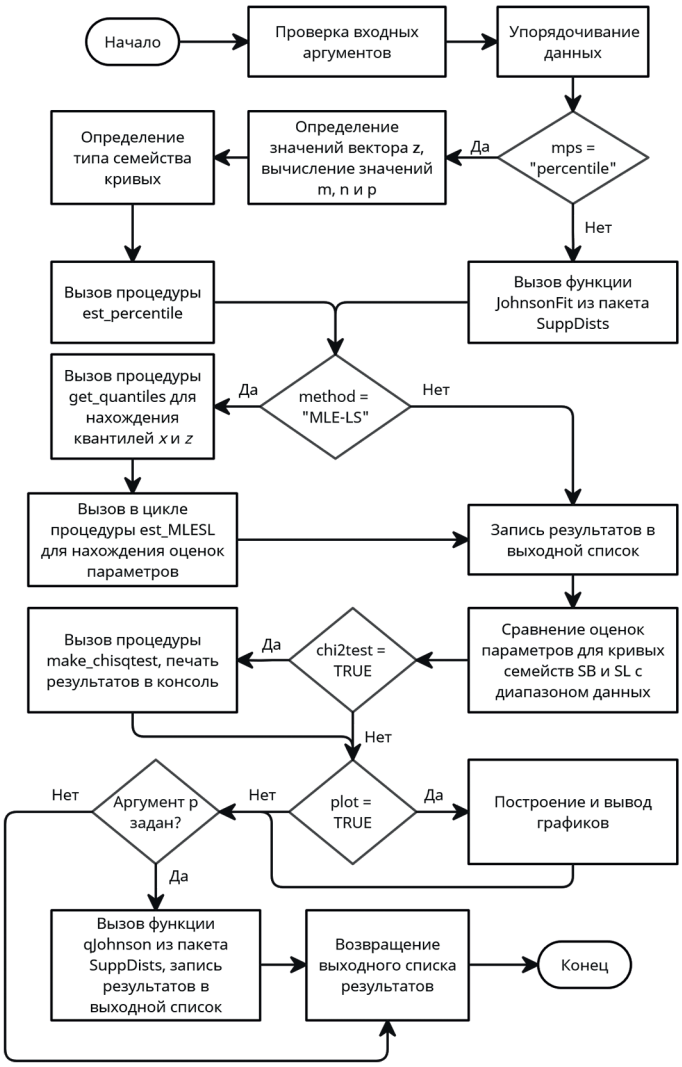

Рис. 1. Общий алгоритм работы функции fit_Johnson

осуществляется объединение внешних соседних интервалов со значениями ожидаемых частот меньше пяти. Кроме того, происходит корректировка числа степеней свободы статистики критерия с учетом связей, налагаемых найденными оценками параметров кривой и, как следствие, корректировка получаемого p – значения;

get_parms_SuppDists — вспомогательная функция для представления результатов оценки в виде, пригодном для вызова функций пакета «SuppDists». Поскольку в функциях из пакета «SuppDists» параметры кривых из системы Джонсона обозначены по-другому и указываются в ином порядке, то данная процедура просто меняет соответствующим образом наименования параме-

тров и располагает их в требуемого порядке для вызова соответствующих функций из данного пакета.

При значениях аргументов msp = "quant" (или "moment") и method = "MLE-LS" для определения семейства кривых Джонсона и получения начальных значений оценок $\hat{\varepsilon}$ и $\hat{\lambda}$ используется функция JohnsonFit из пакета «SuppDists». При значении аргумента plot = TRUE для построения кривой функции плотности (pdf) подобранной кривой используется функция dJohnson, а для построения кривой функции распределения (cdf) — функция pJohnson из этого же пакета. Наконец, если требуется получить процентили подобранной кривой (задан аргумент p), то для этого из пакета «SuppDists» вызывается функция qJohnson.

Поскольку для семейства кривых S_B параметры ε и λ связаны с границами области возможных значений исследуемой случайной величины, то при выборе данного семейства кривых для аппроксимации данных найденные оценки указанных параметров используются для нахождения нижней и верхней границ области возможных значений, с которыми затем сравниваются минимальное и максимальное значения выборочных данных. Если наименьшее выборочное значение оказывается меньше полученной нижней границы области возможных значений и/или наибольшее выборочное значение оказывается больше верхней границы, то выводится предупреждение о том, что полученная аппроксимация может быть неудовлетворительной. Кроме того, при заданных аргументах plot = TRUE или chi2test = TRUE, подобные ситуации часто могут приводить к ошибкам, возвращаемым соответствующими функциями пакета «SuppDists». Аналогичные ситуации возникают и в том случае, если при выбранном семействе кривых S_L оценка параметра ε оказывается больше выборочного минимума.

4. Примеры применения описываемой процедуры

В качестве демонстрации работы описываемой процедуры рассмотрим ее применение к данным из [2] (стр. 353, Задача 170). Вначале выполним подбор кривой с настройками по умолчанию, для этого достаточно подать на вход функции только вектор данных.

```
> x <– c( 1.0, 2.0, 2.5, 3.0, 3.5, 4.2, 4.6, 4.9, 5.0, 5.0,
+       5.1, 5.2, 5.3, 5.3, 5.5, 5.6, 5.8, 6.0, 6.1, 6.1,
+       6.3, 6.5, 6.6, 6.7, 6.8, 7.0, 7.0, 7.1, 7.1, 7.3,
+       7.5, 7.5, 7.9, 7.9, 8.0, 8.1, 8.1, 8.3, 8.4, 8.5,
+       8.7, 8.7, 8.8, 8.8, 8.9, 9.0, 9.2, 9.4, 9.5, 9.5,
+       9.6, 9.7, 9.8, 9.9, 10.1, 10.2, 10.2, 10.2, 10.5, 10.6,
+      10.8, 10.9, 10.9, 11.0, 11.0, 11.2, 11.3, 11.5, 11.7, 11.8,
+      11.9, 11.9, 12.0, 12.3, 12.4, 12.5, 12.7, 12.7, 13.0, 13.2,
+      13.5, 13.9, 14.0, 14.1, 14.2, 14,3, 14.4, 14.5, 14.8, 15.2,
+      15.3, 15.9, 16.0, 16.5, 17.0, 17.5, 18.0, 21.2, 22.3, 26.1)
```

```
> fit_Johnson(x) # настройки по умолчанию

$type
[1] "SB"

$coefficients
    eta   gamma  lambda epsilon
 1.2142  0.7073 24.9000  0.5273

$RSS
[1] 32.76

$method
[1] "MLE-Least Square Approach"

$percentiles
[1] NA

Warning message:
In fit_Johnson(x) :
  Оценки границ области возможных значений не соответствуют результатам наблюдений.
  Подбор кривой может быть неудовлетворительным.
```

Из полученных результатов видно, что подобранная кривая принадлежит семейству S_B с оценками параметров, найденными методом MLE-LS. Однако предупреждающее сообщение указывает на то, что подбор может быть неудовлетворительным из-за несоответствия оценок параметров области возможных значений исследуемой величины. Для улучшения качества подбора можно изменить значение аргумента z или alpha, например, задав аргумент alpha = 0.03 (что примерно соответствует значению аргумента z = 0.627) и, включив проверку качества подбора, получим:

```
> fit_Johnson(x, alpha = 0.03, chi2test = TRUE, plot = TRUE)

    Chi-squared test for given probabilities
data:  Observed Frequencies
X-squared = 0.92, df = 1, p-value = 0.3

$type
[1] "SU"

$coefficients
    eta   gamma  lambda epsilon
  2.884  –2.092   9.292   2.096

$RSS
[1] 7.672

$method
[1] "MLE-Least Square Approach"
```

В результате подобранная кривая уже принадлежит семейству S_U, данные проверки по критерию хи-квадрат Пирсона сви-

детельствуют о достаточно хорошей аппроксимации, что также видно из полученных графиков (рис. 2). Отметим, что для рассматриваемого примера в [2] было подобрано распределение из семейства S_L, однако качество подбора не оценивалось.

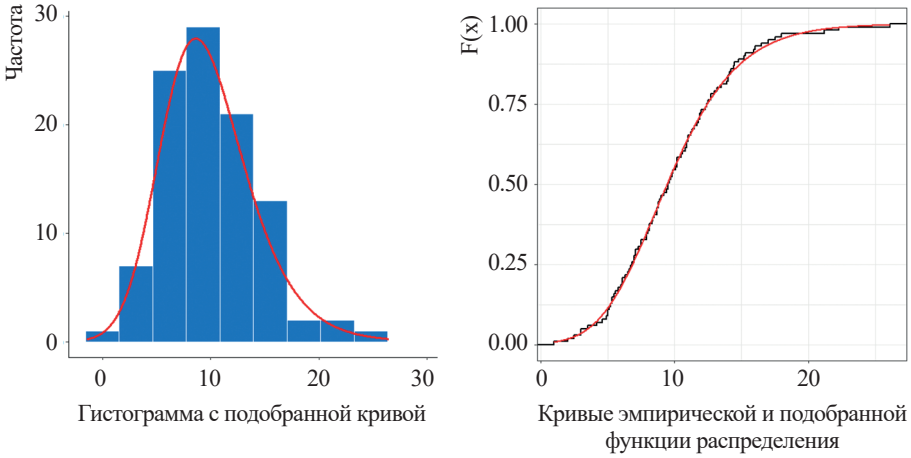

Рис. 2. Графическое представление результатов аппроксимации

В качестве альтернативы можно выполнить подбор кривой с использованием метода процентилей, указав соответствующее значение аргумента method. Из приведенных ниже результатов видно, что при использовании данного метода также подбирается кривая из семейства S_U, но с несколько другими значениями оценок параметров.

```
> fit_Johnson(x, alpha = 0.03, method = "percentile", chi2test = TRUE)

   Chi-squared test for given probabilities

data:  Observed Frequencies
X-squared = 1, df = 1, p-value = 0.3

$type
[1] "SU"

$coefficients
   eta   gamma  lambda epsilon
 2.741 –1.987   8.858   2.423

$RSS
[1] NA

$method
[1] "Percentile Matching"
```

Другой вариант — при заданном по умолчанию значении аргумента method изменить значение аргумента msp, выбрав другой

метод определения семейства кривых и начальных значений оценок параметров. Например, ниже приведен вывод при использовании в качестве начальных значений оценок, найденных методом моментов — видно, что результаты сопоставимы с полученными ранее. (Отметим, что применение метода квантилей приведет к выбору семейства кривых S_L как в [2], но с худшим качеством подбора).

```
> fit_Johnson(x, msp = "moment", chi2test = TRUE)

   Chi-squared test for given probabilities

data:  Observed Frequencies
X-squared = 0.96, df = 1, p-value = 0.3

$type
[1] "SU"

$coefficients
   eta   gamma  lambda epsilon
 2.553 –1.476   8.838   4.068

$RSS
[1] 7.41

$method
[1] "MLE-Least Square Approach"
```

При необходимости, с помощью аргумента p можно задать вектор уровней процентилей подобранной кривой, которые представляют интерес — ниже приведен пример получения процентилей уровней 5 %, 10 %, 90 % и 95 % для подобранной кривой. Наконец, результат подбора можно записать в переменную. В этом случае получить значения оценок параметров подобранной кривой можно с помощью стандартной функции-экстрактора coef, а остальные данные — с помощью оператора $.

```
> fit <- fit_Johnson(x, alpha = 0.03, p = c(0.05, 0.10, 0.90, 0.95))
> coef(fit)

   eta   gamma  lambda epsilon
 2.884 –2.092   9.292   2.096

> fit$type

[1] "SU"

> fit$percentiles

    5%    10%    90%    95%
 3.543  4.742 15.621 17.801
```

В качестве другого примера воспользуемся набором данных «mpg» из библиотеки «ggplot2» и подберем кривую из систе-

мы Джонсона к данным о расходе топлива в городе (переменная «cty»). Применение функции с параметрами по умолчанию приведет к появлению предупреждения о возможном неудовлетворительном подборе. При этом попытка проверки качества подбора графически или по критерию хи-квадрат приведет к появлению ошибки по указанной причине, возвращаемой соответствующей функцией из пакета «SuppDists».

```
> library(ggplot2)
> data(mpg)
> MPG <- mpg$cty

> fit <- fit_Johnson(MPG)

Warning messages:
1: In log(y/(1 – y)) : NaNs produced
2: In log(y/(1 – y)) : NaNs produced
3: In log(y/(1 – y)) : NaNs produced
4: In fit_Johnson(MPG) :
  Оценки границ области возможных значений не соответствуют результатам наблюдений.
  Подбор кривой может быть неудовлетворительным.

> coef(fit)

    eta   gamma  lambda epsilon
 0.9673  0.3054 19.2369  8.4315

> fit$type

[1] "SB"
```

Как и в предыдущем примере, можно попытаться изменить значение по умолчанию аргумента z или задать аргумент alpha, но более гибкий подход состоит в задании в качестве аргумента z вектора равноудаленных значений квантилей стандартного нормального распределения, что позволяет обеспечить наилучший подбор в заданной области, например, на правом хвосте распределения.

```
> fit <- fit_Johnson(MPG, z = c(-2.1, 0, 2.1, 4.2),
+                    chi2test = TRUE, plot = TRUE)

    Chi-squared test for given probabilities

data:  Observed Frequencies
X-squared = 24, df = 5, p-value = 3e-04

> coef(fit)
    eta   gamma  lambda epsilon
  1.846   1.131  35.067   4.190

> fit$type

[1] "SB"
```

Несмотря на то, что в этот раз предупреждения не генерируются и визуальное сравнение подобранной кривой с данными выглядит вполне удовлетворительным, данные проверки по критерию хи-квадрат указывают на то, что результат подбора может оказаться не приемлемым. В этом случае может потребоваться подбор эмпирической кривой из других систем, например Пирсона.

Заключение

Язык программирования R находит все более широкое применение в различных областях деятельности. На практике часто встречаются задачи, связанные с аппроксимацией собранных данных каким-либо эмпирическим распределением из различных систем, в частности, из системы кривых Джонсона. Описанная процедура может рассматриваться как полезное дополнение к уже существующей функции из пакета «SuppDists», реализуя два других метода оценки параметров аппроксимирующей кривой. Кроме того, предлагаемая процедура позволяет гибко настраивать подбор, изменяя соответствующие входные параметры, а также оценивать качество подбора как визуально, так и с использованием критерия согласия хи-квадрат Пирсона.

Список использованной литературы

1. Хан Г. Статистические модели в инженерных задачах / Г. Хан, С. Шапиро. — Москва : Мир, 1969. — 396 с.

2. Кобзарь А.И. Прикладная математическая статистика : для инженеров и науч. работников / А.И. Кобзарь. — Москва : Физматлит, 2006. — 816 с.

3. Johnson N.L. Systems of frequency curves generated by methods of translation / N.L. Johnson // Biometrika. — 1949. — Vol. 36, no. 1/2. — P. 149–176.

4. Johnson N.L. Tables to facilitate fitting S_U curves / N.L. Johnson // Biometrika. — 1965. — Vol. 52, no. 3/4. — P. 547–558.

5. Elderton W.P. Systems of Frequency Curves / W.P. Elderton, N.L. Johnson, — Cambridge : University Press, 1969. — 224 p.

6. Hill I.D. Algorithm AS 99: Fitting Johnson curves by moments / I.D. Hill, R. Hill, R.L. Holder // Journal of the Royal Statistical Society. Series C (Applied Statistics). — 1976. — Vol. 25, no. 2. — P. 180–189.

7. Slifker J.F. The Johnson system: selection and parameter estimation / J.F. Slifker, S.S. Shapiro // Technometrics. — 1980. — Vol. 22, no. 2. — P. 239–246.

8. Wheeler R.E. Quantile estimators of Johnson Curve parameters / R.E. Wheeler // Biometrika. — 1980. — Vol. 67, no. 3. — P. 725–728.

9. George F. Estimation of parameters of Johnson's system of distributions / F. George, K.M. Ramachandran. — DOI 10.22237/jmasm/1320120480 // Journal of Modern Applied Statistical Methods. — 2011. — Vol. 10, iss. 2. — Art. 9.

10. Wheeler B. SuppDists: Supplementary Distributions. R package version 1.1-9.7 / B. Wheeler // CRAN — Package SuppDists. — 2022. — January 3. — URL: https://CRAN.R-project.org/package=SuppDists.

References

1. Hahn G.J., Shapiro S.S. *Statistical Models in Engineering*. New York, Willey, 1967. 355 p. (Russ. ed.: Hahn G.J., Shapiro S.S. *Statistical Models in Engineering*. Moscow, Mir Publ., 1969. 396 p.).

2. Kobzar A.I. *Applied Mathematical Statistics*. Moscow, Fizmatlit Publ., 2006. 816 p.

3. Johnson N.L. Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, 1949, vol. 36, no. 1/2, pp. 149–176.

4. Johnson N.L. Tables to Facilitate fitting S_U Curves. *Biometrika*, 1965, vol. 52, no. 3/4, pp. 547–558.

5. Elderton W.P., Johnson N.L. *Systems of Frequency Curves*. Cambridge University Press, 1969. 224 p.

6. Hill I.D., Hill R., Holder R.L. Algorithm AS 99: Fitting Johnson Curves by Moments. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1976, vol. 25, no. 2, pp. 180–189.

7. Slifker J.F., Shapiro S.S. The Johnson System: Selection and Parameter Estimation. *Technometrics*, 1980, vol. 22, no. 2, pp. 239–246.

8. Wheeler R.E. Quantile Estimators of Johnson Curve Parameters. *Biometrika*, 1980, vol. 67, no. 3, pp. 725–728.

9. George F., K.M. Ramachandran. Estimation of Parameters of Johnson's System of Distributions. *Journal of Modern Applied Statistical Methods*. 2011, vol. 10, iss. 2, article 9. DOI: 10.22237/jmasm/1320120480.

10. Wheeler B. SuppDists: Supplementary Distributions. R package version 1.1-9.7. *CRAN — Package SuppDists,* 2022, January 3. Available at: https://CRAN.R-project.org/package=SuppDists.

Информация об авторах

Борбаць Николай Михайлович — кандидат технических наук, доцент, доцент кафедры «Управление качеством, стандартизация и метрология», Брянский государственный технический университет, г. Брянск, Российская Федерация, e-mail: borbact@mail.ru.

Школина Татьяна Викторовна — кандидат технических наук, доцент, доцент кафедры «Информатика и программное обеспечение», Брянский государственный технический университет, г. Брянск, Российская Федерация, e-mail: shkolina.tv@yandex.ru.

Information about the Authors

Nikolay M. Borbats — PhD in Technical Sciences, Associate Professor, Department of Quality Management, Standardization and Metrology, Bryansk State Technical University, Bryansk, Russian Federation, e-mail: borbact@mail.ru.

Tatiana V. Shkolina — PhD in Technical Sciences, Associate Professor, Department of Computer Science and Software, Bryansk State Technical University, Bryansk, Russian Federation, e-mail: shkolina@yandex.ru.

Вклад авторов

Все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the Authors

The authors contributed equally to this article. The authors declare no conflicts of interests.

Для цитирования

Борбаць Н.М. Процедура подбора кривой из системы Джонсона методами процентилей и максимального правдоподобия — наименьших квадратов в R / Н.М. Борбаць, Т.В. Школина. — DOI 10.17150/2713-1734.2023.5(4).476-493. — EDN AGZLFU // System Analysis & Mathematical Modeling. — 2023. — Т. 5, № 4. — С. 476–493.

For Citation

Borbats N.M., Shkolina T.V. The Procedure for Selecting a Curve from the Johnson System by Percentile Matching and Maximum Likelihood and Least Squares Approaches in R. *System Analysis & Mathematical Modeling*, 2023, vol. 5, no. 4, pp. 476–493. (In Russian). EDN: AGZLFU. DOI: 10.17150/2713-1734.2023.5(4).476-493.