DOI 10.33278/SAE-2018.rus.181-184
УДК 519.254
JEL C12, C49

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ С ИСПОЛЬЗОВАНИЕМ СРЕДСТВ ВИЗУАЛИЗАЦИИ В СРЕДЕ RSTUDIO

Ирина Евгеньевна Денежкина (ORCID 0000-0002-5225-549X) [1],
Сергей Алексеевич Зададаев (ORCID 0000-0003-1329-4012) [2]

[1,2]ФГОБУ ВО «Финансовый университет при Правительстве РФ»

Аннотация. Исследование посвящено разработке и обоснованию средств визуализации критериев принятия решений в задачах проверки статистических гипотез на заданный закон распределения (в частности, проверка нормальности). Рассмотрен общий имитационный подход к графическому построению зоны некритической области на языке программирования R, пригодный для реализации любых критериев (Колмогорова-Смирнова, Пирсона, и др.). Текст статьи содержит рабочие скрипты на R и, полученные с их помощью, графические иллюстрации.

Ключевые слова: визуализация на языке R, критерии нормальности, проверка статистических гипотез.

Abstract. The research is devoted to the development and justification of visualization tools for decision-making criteria in the problems of testing statistical hypotheses for a given distribution law (in particular, the normality test). A General simulation approach to the graphical construction of the non-critical area zone in the programming language R, suitable for the implementation of any criteria (Kolmogorov-Smirnov, Pearson, etc.) is considered. The text of the article contains working scripts on R and graphic illustrations obtained with their help.

Keywords: visualization in R, tests of statistical hypotheses, the criteria of normality.

Введение

Проверка статистических гипотез является неотъемлемой частью анализа данных. Несмотря на огромное количество существующих методов, эта задача неизменно требует новых подходов, соответствующих современному развитию вычислительной техники и соответствующих технологий. Использование различных критериев при проверке статистических гипотез может быть сделано гораздо более комфортным для пользователя. В работе описывается построение графической модели, визуализирующей анализ соответствия исследуемой выборки заданному закону распределения. Приводится решение этой задачи на языке статистического анализа R в среде RStudio. При стандартном подходе ориентируясь только на значение Pvalue по отношению к выбранному уровню значимости, мы не учитываем ошибку второго рода. Имея перед собой графическое представление характер поведения подобных исследуемой выборок, можно более обоснованно заключить соответствует ли значение полученного Pvalue предположению о справедливости нулевой или отклонения в распределении, действительно, имеют место.

1. Визуализация коридора надежности принятия решения

В настоящее время язык R еще не получил достаточного распространения, хотя является одним из самых современных средств получения результатов статистических исследований. Все приводимые в работе коды на языке R универсальны и могу быть запущены с любого компьютера, на котором установлен язык R и удобная интерфейсная оболочка RStudio. О том, как это сделать подробно указано в [1].

Для иллюстрации предлагаемого метода библиотека «nortest», которая позволяет производить по данным вариационного ряда проверку нормальности распределения согласно критерию Колмогорова-Смирнова (К-С) в модификации Лил-

лиефорса [2], [3]. Этот критерий и распределение могут быть заменены на любые другие, оформленные в виде соответствующей процедуры на языке R.

Сгенерируем средствами языка R случайную выборку объема 1000 из нормального распределения $N(4;1)$, эта выборка и будет исследуемой. Применим критерий К-С (см. рисунок 1).

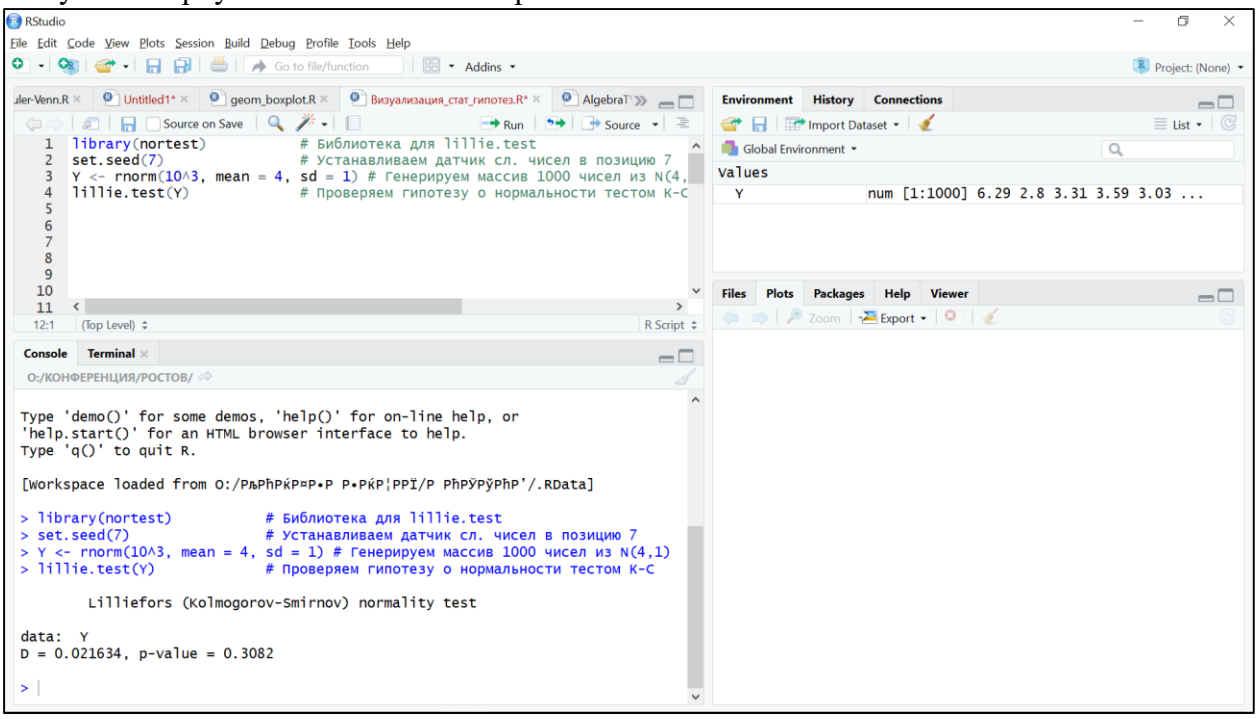

Рис. 1 **Рабочее окно RStudio с результатом применения критерия Колмогорова-Смирнова к сгенерированной выборке**

Заметим, что в левом нижнем окне RStudio (его называют *консолью*) выведен отчет о проверке гипотезы о соответствии нашей выборки нормальному распределению, получено значение Pvalue = 0.3082.

Нашей целью является визуализация того, насколько исследуемый вариационный ряд Y соответствует типичными выборками того же объема из предполагаемого согласно нулевой гипотезе нормального распределения. На практике мы не знаем заранее параметры распределения Y, поэтому, будем использовать для их оценки метод моментов.

Построим кривую плотности вероятности нормального распределения с оцененными параметрами $\mathrm{N}(\mathrm{mean}(\mathrm{Y}), \mathrm{sd}(\mathrm{Y}))$, которое реализует нулевую гипотезу на интервале $[\min_i Y;\ \max_i Y]$. В заголовке графика укажем вычисленное значение Pvalue, округленное до 4 знаков, вместе с заданным уровнем значимости Alpha = 0.05:

```
Alpha <- 0.05
t <- seq(min(Y), max(Y), length = 1000)
plot(t, dnorm(t, mean(Y), sd(Y)), type = "l", lwd = 2,
ylim = c(0, max(dnorm(t, mean(Y), sd(Y))) + 0.1),
main = paste("Alpha = ", Alpha, "; Pvalue = ", round(lillie.test(Y)$p.value, 4)))
abline(v = round(min(Y)) : round(max(Y)), h = seq(0, max(dnorm(t, mean(Y), sd(Y))) + 0.1, 0.1),
lty = 2, col = "gray60")
```

Далее сгенерируем 1000 выборок из $\mathrm{N}(\mathrm{mean}(\mathrm{Y}), \mathrm{sd}(\mathrm{Y}))$ того же объема, что и исследуемая выборка: length(Y). Для каждой из них будем проверять условие: есть ли основания отвергнуть нулевую гипотезу на заданном уровне значимости. Для тех выборок, для которых Pvalue окажется не

меньше Alpha, будем наносить серым цветом график эмпирической функции плотности распределения на предыдущий рисунок.

В совокупности, отобранные из тысячи кривые графически образуют коридор заданной надежности (1-Alpha), попадание в который иллюстрирует нам склонность к выполнению критерия К-С.

Приведенный ниже код выстраивает на предыдущем рисунке визуальный коридор надежности, естественно временно затерев собой теоретическую функцию плотности (см. рисунок 3):

```
for (i in 1:10^3) {X <- rnorm(length(Y), mean(Y), sd(Y))
if (lillie.test(X)$p.value >= Alpha) {lines(density(X), lwd = 1, col = "gray80", type = "l")}}
```

Теперь нанесем на этот график эмпирическую функцию плотности исследуемого вариационного ряда Y (сплошной линией), а также теоретическую плотность вероятности для нулевой гипотезы (пунктиром) (см. рисунок 4):

```
lines(t, dnorm(t, mean(Y), sd(Y)), type = "l", lwd = 1, pch = 19, col = "black", lty = "33")
lines(density(Y), lwd = 2, col = "gray30", type = "l")
```

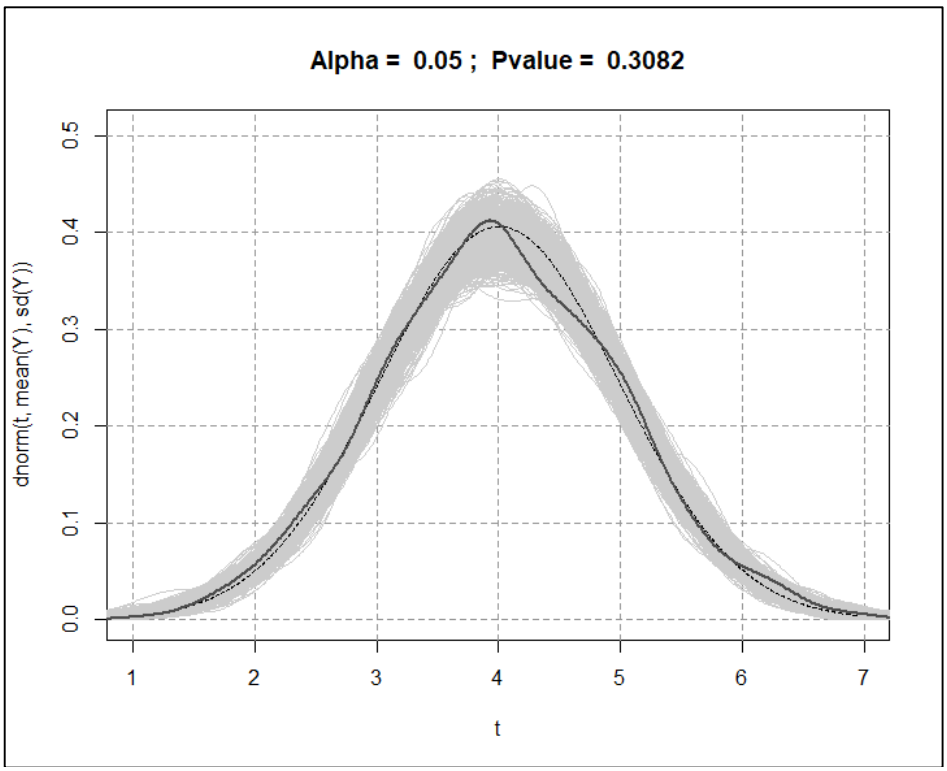

Рис. 2 **Попадание эмпирической плотности вероятности в доверительный коридор**

РАСШИРЕННЫЙ АНАЛИЗ ОТКЛОНЕНИЯ НУЛЕВОЙ ГИПОТЕЗЫ

Полученный рисунок вполне соответствует нашим представлениям о хорошем согласовании наблюдений нулевой гипотезе. Здесь и Pvalue превышает уровень значимости, и кривая находится в полученном коридоре.

Теперь исследуемый вариационный ряд так, чтобы он несколько не соответствовал нулевой гипотезе: сгенерируем 300 случайных значений из распределения Стьюдента: $Y \sim t(3.8)$. После написания такой процедуры на R и построения графиков, аналогичных полученных выше, получим картину, представленную на рис.3.

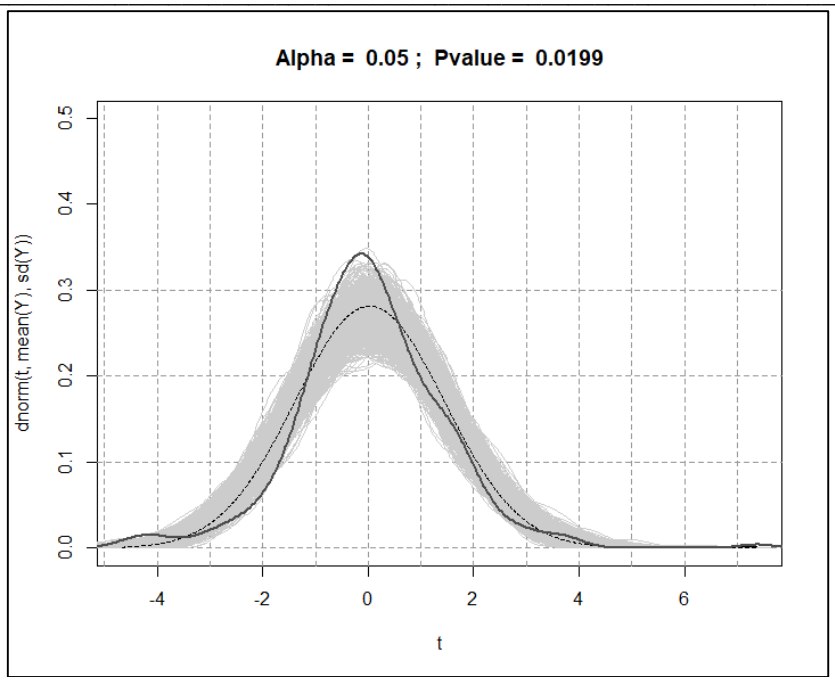

Рис. 3 **Отклонение эмпирической плотности вероятности от доверительного коридора нулевой гипотезы**

На рисунке 3 можно отметить, что небольшое отклонение эмпирической кривой от доверительного коридора плотности вероятности хорошо видно на графике и согласуется со значением Pvalue, которое хоть и незначительно, но все-таки для этого случая уже становится меньше уровня значимости, что приводит к отклонению нулевой гипотезы о нормальности распределения.

Заметим, что такой визуальный анализ имеет смысл только при одновременном с ним количественном сопоставлении выбранного уровня значимости с точным значением Pvalue. Мы целенаправленно не стали выбирать сильно отличающийся от нормального распределения вариационный ряд, т.к. в таком случае все слишком очевидно: и в графике, и в значении Pvalue.

Заключение

Предложенный подход может быть с легкостью перенесен на другой случай, можно применять любой статистический критерий для проверки соответствия исследуемого вариационного ряда любому заданному распределению. Исследователь, владеющий основами языка R, имеет возможность решать свои собственные задачи.

Кроме того, что построение таких графических коридоров надежности для выбранных критериев несут в себе несколько иную, возможно, неформальную, информацию о соответствии. Здесь возникает некоторое не объясненное интегральное свойство «родства» теоретической и исследуемой плотности, данное нам в графических ощущениях.

Список источников

1. Зададаев С.А. Математика на языке R. Учебник. – М.: Издательство «ПРОМЕТЕЙ», 2018. – 324 с
2. Dallal, G.E. and Wilkinson, L. (1986): An analytic approximation to the distribution of Lilliefors' test for normality. The American Statistician, 40, 294–296
3. Thode Jr., H.C. (2002): Testing for Normality. Marcel Dekker, New York