DOI: 10.33278/SAE-2018.eng.150-152

JEL C12, C49

TESTING STATISTICAL HYPOTHESES WITH THE USE OF VISUALIZATION TOOLS IN R STUDIO

DENEZHKINA IRINA (ORCID 0000-0002-5225-549X)[1],
ZADADAEV SERGEY (ORCID 0000-0003-1329-4012)[1]

[1] Financial University under the Government of the Russian Federation

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.

INTRODUCTION

Statistical hypothesis testing is an integral part of data analysis. Despite the huge number of existing methods, this task always requires new approaches corresponding to the modern development of computer technology and related technologies. The use of different criteria when testing statistical hypotheses can be made much more comfortable for the user. The paper describes the construction of a graphical model that visualizes the analysis of the compliance of the sample with the given distribution law. The solution of this problem in the language of statistical analysis R in the environment of RStudio is given. In the standard approach, focusing only on the value of P-value in relation to the selected significance level, we do not take into account the second kind of error. Having a graphical representation of the behavior of such samples, it is possible to conclude more reasonably whether the value of the P-value obtained corresponds to the assumption of the validity of zero or deviations in the distribution, indeed, take place.

VISUALIZATION OF THE DECISION RELIABILITY CORRIDOR

At present, the R language has not yet been sufficiently disseminated, although it is one of the most modern means of obtaining the results of statistical studies. All cited in the work of the codes in R language is universal and can be run from any computer on which you installed the R language and an excellent interface shell RStudio. How to do this is detailed in [1].

To illustrate the proposed method, the library "nortest", which allows to check the normality of the distribution according to the Kolmogorov-Smirnov criterion (K-C) in the Liliefors modification [2], [3], according to the variational series data. This criterion and distribution can be replaced by any other procedure in the form of the corresponding language.

Let's generate a random sample of volume 1000 from the normal distribution N\ left(4;1\right) by means of R language , this sample will be investigated. Apply the K-C test (see figure 1).

Note that in the lower left window of RStudio (it is called console) there is a report on the verification of the hypothesis about the correspondence of our sample to the normal distribution, the value P-value = 0.3082 is obtained.

Our goal is to visualize how the studied variation series Y corresponds to typical samples of the same volume from the assumed normal distribution according to the null hypothesis. In practice, we do not know in advance the parameters of the distribution of Y, therefore, we will use the method of moments to estimate them.

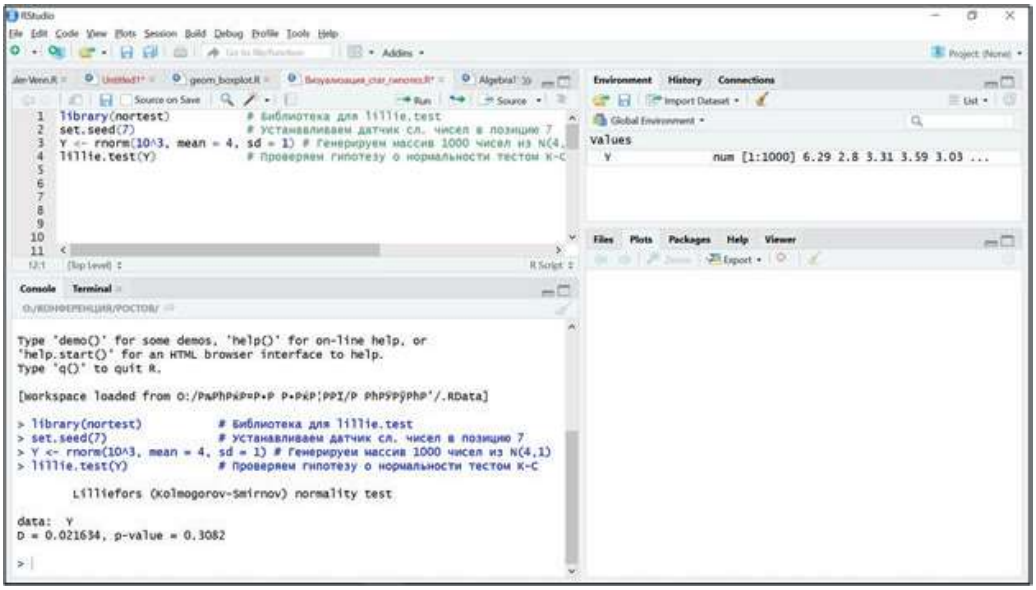

Fig. 1 Rstudio working window with the result of the application Kolmogorov-Smirnov criterion for the generated sample

We construct a probability density curve of the normal distribution with the estimated parameters N\left(mean\left(Y\right),\ sd\left (Y\right)\right), which implements the null hypothesis on the interval [\min\ below I{Y};\ max \below I{Y}] . In the header of the chart, we specify the calculated value of P-value, rounded to 4 characters, together with the specified level of significance 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, “; P-value = “, 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»)
```

Next, generate 1000 samples from N\ left(mean\left(Y\right),\ sd\left(Y\right)\ right)\ of the same size as the sample under study: length (Y). For each of them we will check the condition: is there any reason to reject the null hypothesis at a given level of significance. For those samples for which P-value is not less than Alpha, we will gray the graph of the empirical distribution density function on the previous figure.

Taken together, the curves selected from the thousand graphically form a corridor of specified reliability (1-Alpha), the hit of which illustrates the propensity to fulfill the criterion of K-C.

The following code builds a visual reliability corridor in the previous figure, naturally temporarily overwriting the theoretical density function (see figure 2):

```
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”)
  }
}
```

Now we draw on this graph the empirical function of the density of the studied variation series Y (solid line), as well as the theoretical probability density for the null hypothesis (dotted line) (see figure 2):

```
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”)
```

DOI: 10.33278/SAE-2018.eng.150-152

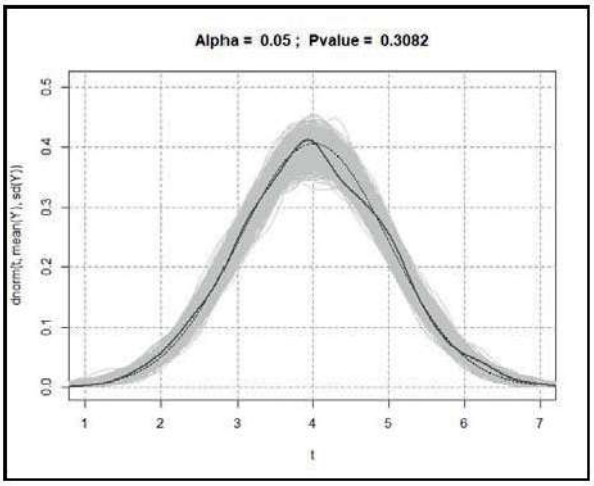

Fig. 2 Getting the empirical probability density in the confidence corridor

ADVANCED ANALYSIS OF THE NULL HYPOTHESIS DEVIATION

The resulting figure is quite consistent with our ideas of a good agreement of observations of the null hypothesis. Here both P-value exceeds the significance level and the curve is in the obtained corridor.

Now the variational series under study do not correspond to the null hypothesis: we generate 300 random values from the student distribution: Y \ ~ \ t\left (3.8\right). After writing such a procedure on R and plotting graphs similar to those obtained above, we get the picture shown in Fig.3.

In figure 3, it can be noted that a small deviation of the empirical curve from the confidence corridor of probability density is clearly seen in the graph and is consistent with the value of P-value, which, though insignificant, but still becomes less important for this case, which leads to a deviation of the null hypothesis about the normality of the distribution

Note that such a visual analysis is meaningful only if simultaneously with it the quantity of material matching the selected level of significance with the exact value of the p-value parameter. Advisedly we haven't chosen the row that dramatically differs from the general one, as in this case both diagram and P-value become too obvious.

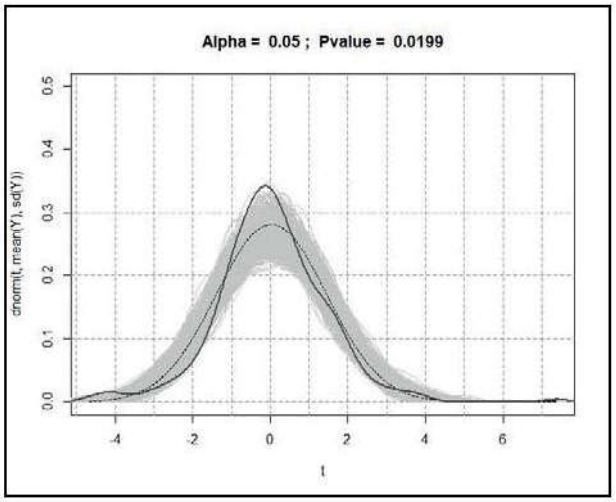

Fig. 3 Deviation of the empirical probability density from the confidence corridor of the null hypothesis

CONCLUSION

The proposed approach can be easily transferred to another case, any statistical criterion can be used to verify the compliance of the studied variational series with any given distribution. The researcher who knows the basics of the R-language has the opportunity to solve their own issu.

In addition, the graphical construction of these corridors reliability for the selected criteria carry a different, perhaps informal, information about compliance. Here there is a not explained by the integral property of the "relationship" theoretical and IP-follow the density given in the graphic sensations.

REFERENCES

1. *Zadadaev S. A.* Math in R. Textbook. – M.: Publishing house «PROMETEY», 2018. - 324

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