International Journal of Economics and Business Administration
Volume VII, Special Issue 2, 2019

pp. 289-296

Visualization of the Decision Criteria in Testing Statistical Hypotheses on Programming in R (Rstudio)

Submitted 20/06/19, 1st revision 27/08/19, 2nd revision 30/10/19, accepted 20/12/19

Bayuk O.A.,[1] Denezhkina I.E.,[2] Zadadaev S.A.[3]

Abstract:

Purpose: *The case study addresses a development and justification of approaches to visualization of decision criteria in the problems of testing statistical hypotheses for a given distribution law (specifically, checking the distribution normality).*

Design/Methodology/Approach: *The study describes a construction of graphical model that visualizes an application of criteria when testing statistical hypotheses for compliance with a given distribution law. This problem is solved in the language of statistical analysis R in the RStudio environment. Using the standard approach and focusing only on the P-value in relation to the chosen level of significance, the researcher cannot take into account the error of the second kind. However, analyzing the graphical representation of the behavior of the sample under study, one can conclude whether the value of the obtained P-value corresponds to the real assumption that the sample corresponds to a given distribution law.*

Findings: *The research case proposes a new approach to testing statistical hypotheses on the compliance of a sample with a given distribution law, using visualization tools and allowing a researcher having even a little experience with the R language to solve applied problems.*

Practical implications: *The approach does not require an in-depth knowledge of mathematics and programming which can be used by experts in various fields of knowledge to successfully solve applied problems. The text of the article contains working scripts in the language R and graphical illustrations obtained with their help.*

Originality/Value: *The main contribution of this study is to expand the variety of methods for testing statistical hypotheses. The proposed method extends the set of statistical problems successfully solved by means of R.*

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

JEL Code: *C12, C15, I10, I12.*

Paper type: *Case study: Statistics.*

[1] *Associate Professor of the Department of Data Analysis, Decision-Making and Financial Technology of Financial University under the Government of the Russian Federation, Moscow, OABayuk@fa.ru*

[2] *Associate Professor of the Department of Data Analysis, Decision-Making and Financial Technology of Financial University under the Government of the Russian Federation, Moscow, IEDenezhkina@fa.ru*

[3] *Associate Professor of the Department of Data Analysis, Decision-Making and Financial Technology of Financial University under the Government of the Russian Federation, Moscow, SAZadadaev@fa.ru*

1. Introduction

Competition among the statistical analysis software presented on the market contributed to the development of object-oriented programming (for mathematicians and programmers) and functional programming (for applied statisticians). The programming languages R and Python, as well as specialized application software libraries created for them, are gaining popularity. These programming languages successfully compete with such well-known statistical programs like SPSS, STATISTICA and EXCEL. Among the most successful moves recently made in this direction the development of Microsoft Azure and Russian Loginom should be mentioned. It promotes the use of various tools of object-oriented design, accessible not only and not so much to for professional mathematicians and programmers, but also to for specialists in applied fields.

In this case study we describe creating a graphical model that visualizes the application of criteria for testing statistical hypotheses for compliance with a given distribution law realized by means of the language of statistical analysis R in the R-Studio environment. Using the standard approach and focusing only on the P-value in relation to the chosen level of significance, we do not take into account the error of the second kind. However, facing graphically represented sample behavior, we could be more certain in concluding whether the obtained P-value corresponds to the null hypothesis about the distribution law or rather significant deviations from the assumed distribution occur.

2. Statistical Analysis Environment R (R-Studio)

Recently, in Russian academic milieu popularity of the R language has only been increasing, but, even though this language is one of the most efficient modern tools to process statistical data and to gain reliable results, it has not been acknowledged to a sufficient degree yet. All quoted codes in the R language are universal and can be run on any computer if only a proper R language environment and a convenient RStudio interface shell are installed. For detailed guidelines see (Zadadaev, 2018, 12-14).

In this case study the library “nortest” is being used, which allows to test the normality of a variational series distribution according to the Lilliefors modification of the Kolmogorov-Smirnov (K-S) (Razali and Yap, 2011; Bayuk *et al.*, 2014). We have chosen this library and criterion just as an example. In practice, the K-S criterion can be replaced by any other, drawn up in the form of an appropriate procedure in R, and for an arbitrary distribution. In order to apply the above mentioned K-S criterion to testing a distribution normality, you need to assign a variable, say Y, as the name of the variation series to study, and then call the lillie.test (Y) command.

Hence, we will use the R language to generate a random sample of a volume of 1000 from the normal distribution N (4; 1) under the name Y, examine this sample for compliance with the specified distribution type, and then apply the K-S criterion of normality. The code for this part of the text will look like this in the RStudio environment:

```
library(nortest)
set.seed(7)
Y <- rnorm(10^3, mean = 4, sd = 1)
lillie.test(Y)
```

In the second line of the code, we set a random number counter (in our case, position 7) so that the results on the reader's computers coincide with those given here.

The reader can do this on his own. As a result, we get a working window of the RStudio program with the code in the R language and the result of applying the Kolmogorov-Smirnov criterion to the generated sample.

Note that in the lower left RStudio window, called the console, the result of testing the hypothesis about the normality of our sample is reported and the Pvalue = 0.3082 is displayed. The pvalue can also be directly obtained with the command
lillie.test (Y) $ p.value.

Our goal now is to visualize how just generated variational series Y corresponds to typical samples of the same volume from the normal distribution as is assumed according to the null hypothesis. Initially that was N (4; 1), however, in more general case, we may not know in advance the distribution parameters Y, and, therefore, we will use one of the possible options for their estimation - the method of moments, i.e. instead of 4, we will use the mean mean (Y), and instead of the standard deviation, the standard deviation sd (Y).

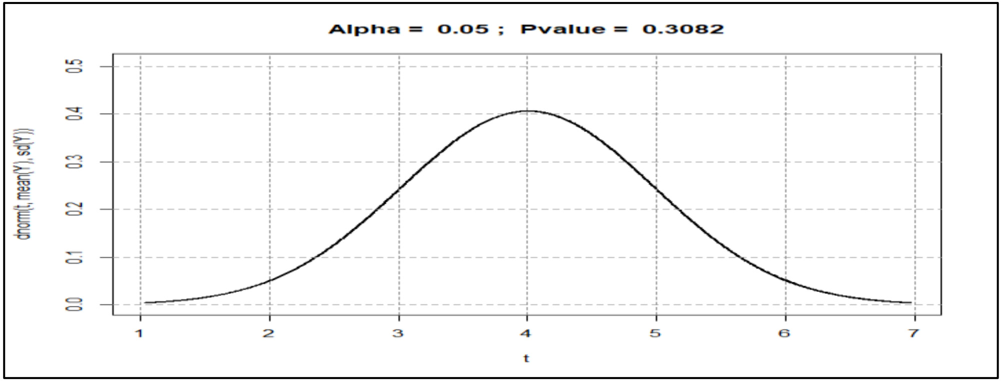
To solve this problem, you need to write a small program, which we will do in the RStudio environment.

3. Visualization of the Corridor of the Reliability Decision Making

Let's consider the previously generated sample Y as a real variational series, the normality of which has to be tested. We draw the probability density curve of the normal distribution with the estimated parameters N (mean (Y), sd (Y)), which implements the null hypothesis on the interval $[\min_i Y;\ \max_i Y]$. In the header of the graph, we indicate the calculated Pvalue, rounded up to 4 characters, along with the given significance level Alpha = 0.05 (Figure 1):

```
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")
```

Figure 1. *The probability density function for the null hypothesis N (mean (Y), sd (Y))*



Further, we will generate 1000 samples of N (mean (Y), sd (Y)) of the same volume as the sample under consideration length (Y). For each one we will check whether there is any reason to reject the null hypothesis at the given level of significance. If the Pvalue is above Alpha, we will add in gray the curve of the empirical density distribution function to the previous graph.

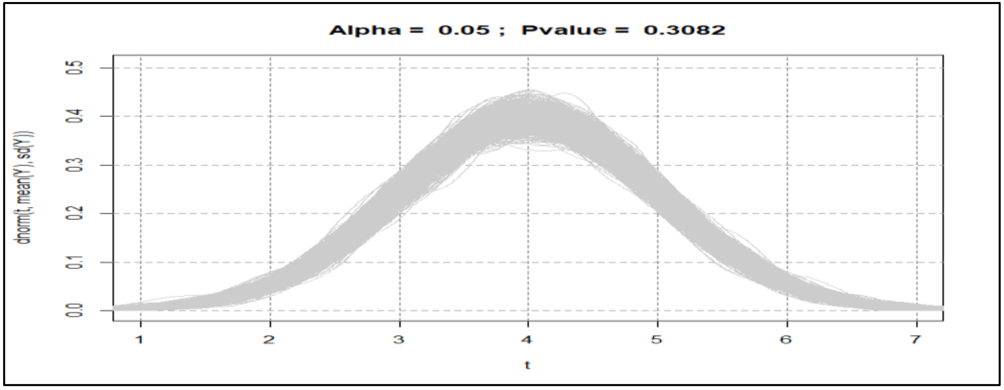
Altogether, the selected curves will graphically form a corridor for the given reliability (1-Alpha). To be within this corridor shows an inclination to meet the K-S criterion. Why only an inclination? We avoid drawing any exact affirmation because the density function may be close to a theoretical curve in the class of one smoothness, but far in the class of another smoothness. The final conclusion also will depend on the choice of the averaging window applied for the empirical probability density function, and the general question whether the sample smoothness corresponds to the normal law remains open.

However, the combined analysis based on both the Pvalue and the graphical representation of the density curve (which may or may not be inside the constructed confidential corridor removes the mentioned difficulties.

The code below builds a visual reliability corridor in the previous figure, naturally temporarily excluding the theoretical density function (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")
```

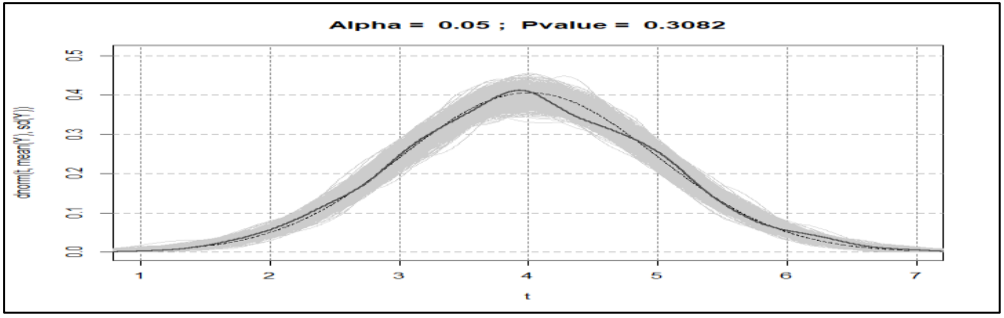
Figure 2. *Confidential corridor consisting of empirical density functions that satisfy the criteria at a given level of significance Alpha*



Now let's plot on this graph the empirical density function of the variational series under study Y (solid line), as well as the theoretical probability density for the null hypothesis (dotted line) (Figure 3):

```
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")
```

Figure 3. *Empirical probability density hit in the confidential corridor null hypothesis at a given level of significance alpha*



4. Analysis of Deviation of the Null Hypothesis

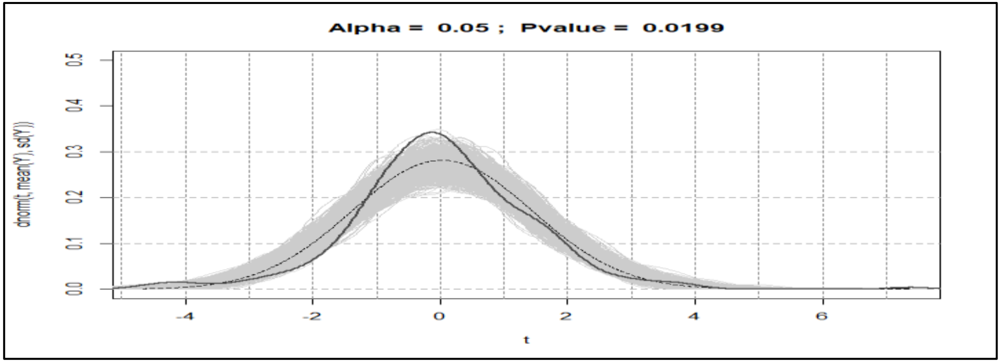
According to the main hypothesis, the sample Y is generated from the normal distribution. The last figure illustrates the compliance of the observations with the null hypothesis. As can be seen from the figure, the Pvalue exceeds the significance level, and the curve is in the resulting corridor.

Now we will add slight modifications to the sample in order to make it deviate from the null hypothesis. For the purpose we'll generate 300 random values from T-distribution t(3.8). For the sake of convenience, we give the procedure code below:

```
set.seed(8)
Y <- rt(300, 3.8)
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, 4, 1))
+ 0.1), main = paste("Alpha = ", Alpha, ";  Pvalue = ", round(lillie.test(Y)$p.value,
4))) abline(v = round(min(Y)):round(max(Y)), h = seq(0, max(dnorm(t, mean(Y),
sd(Y))) + 0.1, 0.1),
lty = 2, col = "gray60")
for (i in 1:10^3) {X <- rnorm(length(Y), mean(Y), sd(Y))
  if (lillie.test(X)$p.value >= Alpha) {lines(density(X), lwd = 1, col = "gray80", type
= "l")}
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")
```

Figure 4 shows the empirical curve does not entirely fit within the confidence corridor of density, and this misfit is in a good correspondence with the Pvalue, which is slightly below the significance level for this case. Therefore, the null hypothesis must be rejected.

Figure 4. *Loss of empirical probability density in the confidential corridor null hypothesis at the given significance level Alpha*

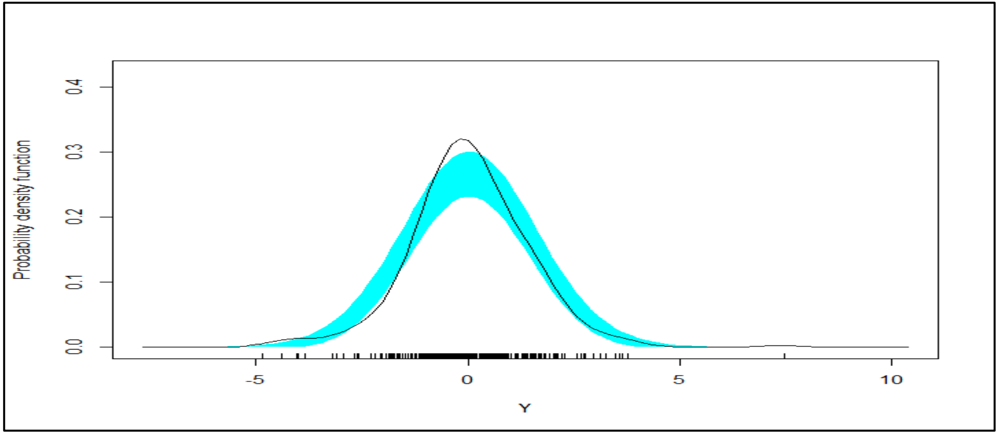


We have to admit that the developed ideas are far from being very original and, even though not too often, could be encountered in some publications on the imitational modeling. For example, in the «sm» library you can solve a similar task of building a visual illustration. It is also true that the method is applicable only in rather few particular cases with fixed significance. The script that calls this procedure for our sample follows:

```
library("sm")
sm.density(Y, model = "Normal")
```

The outlook on the Figure 5 similarly indicates the non-normal distribution of the studied variational series Y.

Figure 5. *Immersion of the graph of the empirical probability density function in the confidence corridor of the null hypothesis at the 0.05 significance level*



Note that such a visual analysis makes sense only if it is completed with a simultaneous quantitative comparison of the chosen level of significance with the exact Pvalue. We deliberately neglect variation series that were too far from the normal distribution, since all becomes obvious in this: both on the graph and in the Pvalue.

5. Conclusion

The proposed approach can be easily expanded for any analogous case; any statistical criterion can be applied in a similar way to verify that the variation series under investigation fits to any given distribution. A researcher, even not familiar enough with the R, could adapt the program codes given in the paper to solving analogous problems with minimal modification. We also note that the construction of such graphical corridors of reliability for the selected criteria provides a new, visual, approach for testing the sample compliance.

This consideration, apparently, cannot in the strict sense of the word be called the pure visualization of the applied criterion. Here there arises some unexplained integral property of the "relationship" of the theoretical and investigated density, given to us in graphic sensations.

References:

Zadadaev, S.A. 2018. Mathematics in R. Textbook. Moscow, Prometey.

Razali, N.M., Yap, B.W. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33.

Bayuk, O.A., Brailov, A.V., Denezhkina, I.E., Zadadaev, S.A. 2014. Decision-making under conditions of comparative uncertainty in financial and economic problems. Moscow, INFRA-M, 106.

Denezhkina, I.E., Zadadaev, S.A. 2016. Probabilistic restoration of the investment attractiveness rating according to the international rating agencies. Scientific works of the Free economic society of Russia, 198, 355-359.

Thode, Jr.H.C. 2002. Testing for Normality. Marcel Dekker, New York.

Denezhkina, I.E., Zadadaev, S.A. 2018. Testing statistical hypotheses with the use of visualization tools in r studio - System analysis in economics. Proceedings of the V International research and practice conference-biennale, 150-152.