The Test of Univariate Normality and Multivariate Normality Based on R Language

Manli Zhang, Rui Chen, Jing Li
TaiShan University, Tai'an 271000, China.

Abstract: The normal distribution is a widely used distribution and has a very important probability distribution in the fields of mathematics, physics and engineering. In natural and social phenomena, a large number of random variables obey or approximately obey the normal distribution, so it is always customary to assume that they are in line with normality when doing data analysis, but whether the assumption is true or not depends on the normality of test. Therefore, the judgment method of normal distribution is also particularly important. This paper proposes corresponding test methods for univariate normality and multivariate normality under the premise of R language.

Keywords: Normal Distribution; Test Method

1. Univariate normality test method

The status and role of the normal distribution in statistics is unquestionable. The normal distribution is also called "normal distribution", also known as Gaussian distribution. Since it was proposed, it has attracted the attention of later generations. In the specific statistical analysis, when we get a sample data how to judge whether it comes from a normal distribution, we will now sort out the feasible methods. In fact, it is to judge the data distribution, which is called the distribution goodness-of-fit test. From the big aspect, we divide it into three categories of methods: image method, skewness and kurtosis method and statistical test method[1].

First of all, let's look at the image method. We mainly judge from the histogram and the QQ graph. The normal distribution curve is bell-shaped, with low ends at both ends, high in the middle, and left and right symmetry. Because the curve is bell-shaped, people often call it a bell shape curve. Based on this characteristic of normal distribution, R language is used to generate four groups of data, normal distribution, exponential distribution, T distribution, and exponential distribution corresponding to the opposite numbers of the data; these four groups of data histograms can be seen from the normal distribution and T The distribution is not easy to distinguish, so how to distinguish this situation? In this case, we use the QQ map to determine.

The full name of the QQ plot is quantile-quantile plot. From the name, we can know that it is determined by quantiles. One quintile is the quantile representing the normal distribution, and the other quintile is the quantile of the actual sample data. , the two are compared to determine whether they come from a normal distribution. If the QQ plot shows the style of a straight line, we judge that it is from a normal distribution, otherwise it is not from a normal distribution. From the QQ plot, we can see that only the QQ plot corresponding to the normally distributed random numbers conforms to this characteristic.

Image method [2] is a relatively rough method, and this method is not suitable for small samples. Skewness-kurtosis and statistical tests can address the limitations of image methods. For the normality test method, a large number of simulation calculations carried out by Okuno Chuichi and others in the 1970s revealed that the "skewness and kurtosis test method" [1] is very effective, that is, the skewness is 0 and the kurtosis is about 3. It is reasonable of. In R language, skewness() and kurtosis() are used to calculate skewness and kurtosis to obtain judgment results.

The following tests, namely Shapiro-Wilks test, Kolmogorov-Smirnov test, Cramer-von Mises test, Anderson-Darling

test, can be used to determine the assumption of normality. These four tests use the functions shapiro.test(), ks.test(), cvm. test(), ad.test() in the R language respectively. We still use the data set 1 of the four sets of data generated at the beginning. After verification, the obtained results are shown in Table 1:

Table 1: Test P value table

DistributionP-value / methodologies	distribution I	distribution II	distribution III	distribution IV
Shapiro-Wilks	0.8131	4.513e-10	1.488e-09	4.513e-10
Kolmogorov-Smirnov	0.9067	0.001676	0.01809	0.001676
Cramer-von Mises	0.5007	7.37e-10	5.184e-08	7.37e-10
Anderson-Darling	0.5258	3.969e-16	3.724e-10	3.969e-16

From the table, we can see that the p-values of distribution 1 are very large for the above four test methods, so accept the null hypothesis that the distribution obeys a normal distribution; the p-values of other distributions are very small, so the null hypothesis is rejected, that is, neither Following a normal distribution, this result is consistent with the facts.

2. Multivariate normal test method

The test method of the univariate normal distribution is introduced above. The test of the multivariate normal distribution should use the test method of the univariate normal distribution and the properties of the multivariate normal distribution. Two methods are introduced here. The first method is based on the marginal distribution of the multivariate normal distribution, that is, the corresponding univariate distribution is also a normal distribution, so it is only necessary to test whether each dimension is univariate normal. Method 2 The correlation coefficients of the multivariate normal distribution are all non-zero, which means that the relationship between the two dimensions has a linear trend. If it is not linear, it does not obey the multivariate normal distribution. Let's take the data set USairpollution that comes with the R language as an example. First draw the QQ diagram of each dimension as shown in Figure 1: You can also further draw the relationship diagram between the two variables to determine whether the linear relationship is in line with the assumption of normality[3].

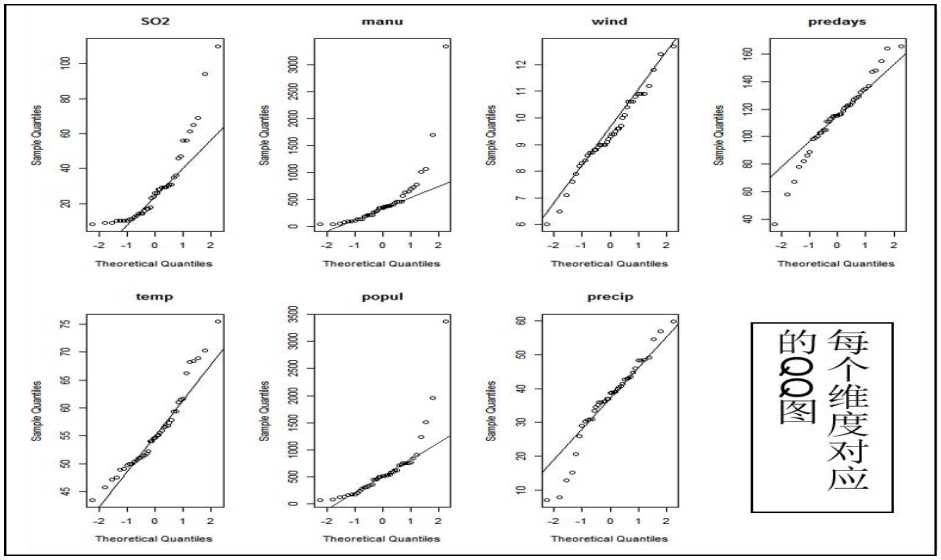

Figure 1: QQ Graph

3. Conclusion

The normality test is a very important topic in statistics. The above-mentioned normality test methods have their own advantages and disadvantages. The image method is very intuitive, but it also has certain limitations, that is, it is highly subjective, and it is not suitable for small samples. Not very suitable. Statistical tests are relatively more rigorous. The most

important basis for the multivariate normality test is the univariate normality test method.

References

[1] Shen HF. Course of Probability Theory and Mathematical Statistics [M]. Beijing: Higher Education Press, 1998.

[2] Wang BH. Graphical method of normality test and its application, Mathematical Statistics and Applied Probability [J].1996, 9(3): 250-255.

[3] Shaoiro SS. An analysis of variance test for normality, Biomerika [J]. 1995(52): 591-611。

About Authors:

Manli Zhang(1985.9-),Female,Han Nationality,Shandong Tai'an people,Lecturer, master.

Rui Chen(1981.04-),Female,,Han Nationality,Shandong Xintai people,Associate Professor， Undergraduate, research direction: mathematics teaching research.

Jing Li(1981.2-), Female, Han Nationality, Shandong Feicheng people, Lecturer, master student, research direction: topology on lattice and fuzzy reasoning.

Fund Project: Taishan College Teaching Reform and Research Project