ICSCCD 2018 p-ISSN 2716-1919 e-ISSN 2716-2907
Volume 1, January 2020 doi:10.31327/icusn-adri.v1i0.1152

Data Normality Detection Using the Graph Method

A Nasrum

University of Sembilanbelas November Kolaka, Indonesia

E-mail: akbar.nasrum@gmail.com

Abstract. The main purpose of this research is to offer a normality test using the graph method. The steps taken by the researcher are conducting a literature study on the normality test of the data. Look for the advantages and disadvantages of each method and compare with other graphics methods such as QQ-Plot and PP-Plot. The last step is to simulate data using the proposed method and then compare the results with other methods and draw conclusions from the results obtained.From the results of the literature study, we found a way to test data only by looking at the graph. The graph that appears when testing the data has three lines. The graph that appears when testing the data has three lines. They are upper, middle and bottom lines.If all data entered in the upper and lower boundary environment, it can be concluded that the data is normally distributed. Using this method, the normal detection of data can be seen only through a graph and the results are more accurate than the Kolmogorov-Smirnov method is some cases.

1. Introduction

In certain studies, especially in the field of education, it is usually required that data must be normally distributed. There are many methods that can be used to test the normality of data.They are *Kolmogorov-Smirnov, Lilliefors, Shapiro-Wilk, Shapiro-Francia, Anderson Darling, Chi-Square, Jarque-Bera, Ryan Joiner, Kuiper, Modified Kuieper, Vasicek, Ajne, Cramer Von Mises and much more.* These methods do not have the same characteristics and strength in detecting the normality of data so that the selection of the right method is key in fulfilling normality requirements [1] .

Some good methods are used for large samples but are very poor at detecting normality if the sample size is small. This error will affect the selection of parametric or nonparametric methods [2]**.** To see which method is the best, a comparison of methods needs to be done. Many researchers compare one method with another method and produce different conclusions.

Yap and Sim measured the strength of eight different normality tests using Monte Carlo simulations [3]. Yazici and Yolacan also used Monte Carlo to distinguish the strength of twelve methods [4]. [5] and [6] each compared four methods and the results obtained varied depending on the method used.

The several methods mentioned above, only Kolmogorov Smirnov has a graphical representation [7]. There are several graphical methods that are often used to test data normality, namely QQ-Plot and PP-PlotGraph methods are rarely used because conclusions that can be obtained depending on one's vision. The distant interpretation of the point towards the line varies for each individual.The problem is that there is no size on the graph that shows to what extent the distribution of points is said to be far from the line. Even the knowledge of the line is still very minimal.

ICSCCD 2018 p-ISSN 2716-1919 e-ISSN 2716-2907
Volume 1, January 2020 doi:10.31327/icusn-adri.v1i0.1152

In this paper, we propose how to detect normal data using graph methods such as QQ-Plot. The process of forming a line that is used as a benchmark for normal distribution of data will also be explained.In order to find out the proximity of the data to the line formed, the confidence interval is also displayed. Thus we can detect normal data only by looking at the graph. If all data points are included in the confidence interval, it can be concluded that the data comes from a population with a normal distribution, whereas data is not normally distributed.

2. Materials and Methods

In this section, we will explain two methods commonly used in data normality tests, namely the Kolmogorov Smirnov test, QQ-Plot. Next will be explained the proposed graph method then compare the results in several ways above.

2.1. Kolmogorov Smirnov Test

The principle of normality test using Kolmogorov-Smirnov is to find the largest deviation from the cumulative distribution function of observation data (empirical) to the theoretical cumulative distribution function. If the maximum deviation is not too large ($D > D_{tab}$), the observation data can be categorized as a normal distribution.

Conversely, if the maximum deviation that is formed is very large ($D > D_{tab}$), the observation data is said to be not normally distributed. The Kolmogorov Smirnov Test Statistics are defined as follows:

$$D = \max_{1 \le i \le n} \left(\left| F(Z_i) - F_{n_{i-1}}(x_i) \right|, \left| F(Z_i) - F_{n_i}(x_i) \right| \right)$$

with $F(Z)$ is the theoretical cumulative distribution function (Normal Standard Z) and $F_n(x)$ is the cumulative distribution function of the observation data [8].

2.2. Normality Test with QQ Plot

Normality test with QQ-plot using the graph method. Another graph method that is almost the same is PP-Plot. The principle is very simple. Data normality is measured based on the proximity of data points on one line which is the expected value of a data.

If the data distribution is near the center line, then the data distribution is normal. The problem is that the proximity of the data on the line is relative. It could be that one researcher said the distribution of data was close to the line but other researchers said different things.Therefore, in the Minitab software, the proximity of the data in the midline can be measured by displaying the confidence interval. If the distribution of data does not come out of this band, it can be ascertained that the data is normally distributed.

In fact, not all software has this capability. In the SPSS the confidence interval line limit cannot be displayed. To measure the proximity of the data to the line, correlation analysis was used. How to use it, the following will be explained.

QQ-Plot is a Scatter Plot between observation data and expected values in a normal distribution. The presentation of the data can be in the form of original data, standardized data (normalized) or a combination of both. The data presented in the scatter diagram has been separated based on the quantiles. The steps in testing data normality using QQ-Plots can be seen in [9].

To achieve the desired results, researchers conducted a literature study of the normality test of data. Look for the advantages and disadvantages of each method and compare with graph methods such as QQ-Plot and PP-Plot. The last step is to simulate the data using the proposed method and then compare the results with other methods and finally draw conclusions on the results obtained.

2.3. Methods Offered

The method offered is actually the development of the QQ plot. Data normality is measured based on the proximity of data points on one line which is the expected value of a data. If the distribution of

ICSCCD 2018 p-ISSN 2716-1919 e-ISSN 2716-2907
Volume 1, January 2020 doi:10.31327/icusn-adri.v1i0.1152

data is close to the line, then the data distribution is normal. It was explained earlier that the problem is that the proximity of the data to the line is relative.

In this paper, the normality test with the proposed graph method displays the confidence interval so that in the picture there are three straight lines that become a benchmark. Look at Figure 1.

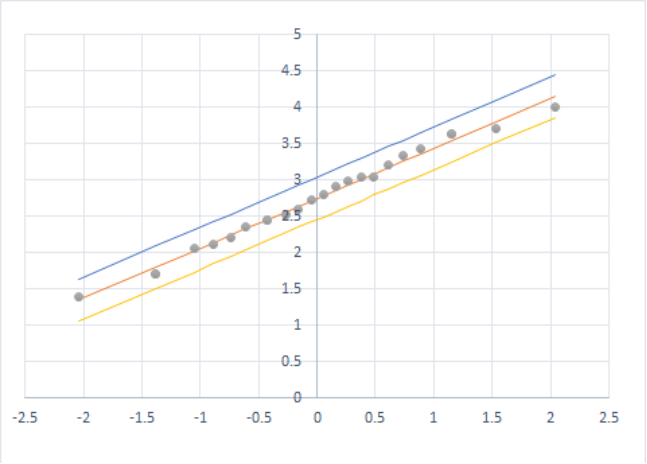

Figure 1. Test for normality using the graph method.

The center line is the expectation value for normally distributed data obtained from the regression results between the expected value of the data and itself. The other two lines are the confidence interval. If the center line is y, then the upper and lower lines are $f(x) = y + t_{\frac{\alpha}{2}(n-1)} \frac{S}{\sqrt{n}}$ and $g(x) = y - t_{\frac{\alpha}{2}(n-1)} \frac{S}{\sqrt{n}}$. For each x, the values of $f(x)$ and $g(x)$ are respectively the lower and the upper limit of the estimation of the expected range.

If all data is included in the band, there is a similarity in the distribution between the data measured by the normal distribution. Conversely, if there is one or more than one data outside the confidence interval, the distribution of measured data deviates from the normal distribution.

3. Result and Discussion

The following sample data will be presented to be tested with two different methods, namely Kolmogorov-Smirnov and the proposed graph method. Next, it will be compared which of the two methods is more sensitive in detecting deviations from the normal distribution. Pay attention to the following data.

Table 1. Sample data

No	1	2	3	4	5	6	7	8	9	10	11
x	65	67	69	70	78	80	85	87	89	90	95
f	2	3	3	3	2	4	3	4	2	3	1

The following details are obtained using the Kolmogorov Smirnov method

Table 2. Kolmogorov-Smirnov statistical calculation procedure manually.

No	x	f	fk	Fk(i)	Fk(i-1)	z	F(z)	a(i)	b(i)
1	65	2	2	0.067	0.000	-1.491	0.068	0.068	0.001
2	67	3	5	0.167	0.067	-1.278	0.101	0.034	0.066
3	69	3	8	0.267	0.167	-1.065	0.143	0.023	0.123

ICSCCD 2018 p-ISSN 2716-1919 e-ISSN 2716-2907
Volume 1, January 2020 doi:10.31327/icusn-adri.v1i0.1152

4	70	3	11	0.367	0.267	-0.959	0.169	0.098	**0.198**
5	78	2	13	0.433	0.367	-0.107	0.458	0.091	0.024
6	80	4	17	0.567	0.433	0.107	0.542	0.109	0.024
7	85	3	20	0.667	0.567	0.639	0.739	0.172	0.072
8	87	4	24	0.800	0.667	0.852	0.803	0.136	0.003
9	89	2	26	0.867	0.800	1.065	0.857	0.057	0.010
10	90	3	29	0.967	0.867	1.172	0.879	0.013	0.087
11	95	1	30	1.000	0.967	1.704	0.956	0.011	0.044
		30							

The value of D = 0.198 is obtained from the results of manual calculations. This value is the same as the following SPSS results.

Table 3. KS test results using SPSS

	Kolmogorov-Smirnov[a]		
	Statistic	**df**	**Sig.**
VAR00001	.198	30	.004

a. Lilliefors Significance Correction

The value of D_{tab} for $\alpha = 0.05$ is 0.242 Because the value of Dhit is smaller than Dtab, it means that deviations from the normal distribution are not significant. So it is concluded that the data is still normally distributed.Another case is if the conclusion is based on corrected P-Value from Lilliefors. From the Table 3 obtained P-Value is 0.004. This value is much smaller than the significance level used, which is 0.05 so that the conclusion of the data is not normally distributed.

What about the normality test using image detection? By using the QQ Plot steps and a few modifications the following results are obtained:

Table 4. Normal test procedure with the graph method.

i	**x**	**j**	**proporsi** (j-0.5)/n	**z**	**expt**	**low**	**upp**
1	65	1.5	0.033	-1.8	61.78	58.28	65.29
2	67	4.0	0.117	-1.2	67.81	64.31	71.32
3	69	7.0	0.217	-0.8	71.64	68.14	75.15
4	70	10.0	0.317	-0.5	74.52	71.02	78.03
5	78	12.5	0.400	-0.3	76.62	73.12	80.13
6	80	15.5	0.500	0.0	79.00	75.49	82.51
7	85	19.0	0.617	0.3	81.79	78.28	85.29
8	87	22.5	0.733	0.6	84.85	81.34	88.35
9	89	25.5	0.833	1.0	88.08	84.58	91.59
10	90	28.0	0.917	1.4	91.98	88.48	95.49
11	95	30.0	0.983	2.1	98.98	95.47	102.48

ICSCCD 2018 p-ISSN 2716-1919 e-ISSN 2716-2907
Volume 1, January 2020 doi:10.31327/icusn-adri.v1i0.1152

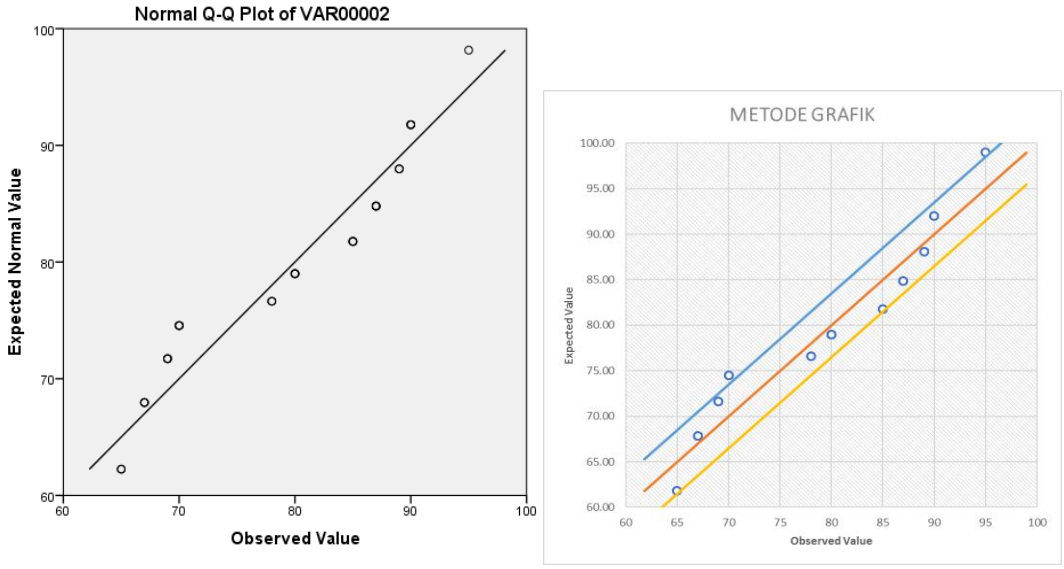

Figure 2. The left-side graph is Q-Q Plot the result of SPSS and the right-side of the result manually using excel.

The above 30 data are grouped according to their quintiles so that they become 11 quantiles with different values. Therefore in the picture, there are only 11 points that represent 11 different values.

The center line is a regression line between the z value and the expectation value x based on the value of the proportions. The equation of the regression line is$y = 79 + 9.38818x$. The top interval line is $f(x) = y + t_{\frac{\alpha}{2}(n-1)} \frac{S}{\sqrt{n}} \approx 82.506 + 9.388x$ and the lower interval is $g(x) = y - t_{\frac{\alpha}{2}(n-1)} \frac{S}{\sqrt{n}} \approx 75.494 + 9.388x$. From the graph, it can be seen that the data in the 4th quantile and the last quantile are outside the confidence interval. So it can be concluded that the samples come from populations that are not normally distributed.

To be more certain whether it is true that the data is outside the line, consider Table 4. The real value of x for the 11th quantile is 95 while the value in the interval of the test is 95.47. This value is above the actual value so that the point is outside the line. Likewise with data in 4th quantile. The real value is 70 while the value on the confidence interval is 71.02. thus the data is outside the confidence interval. Look at Figure 3 below.

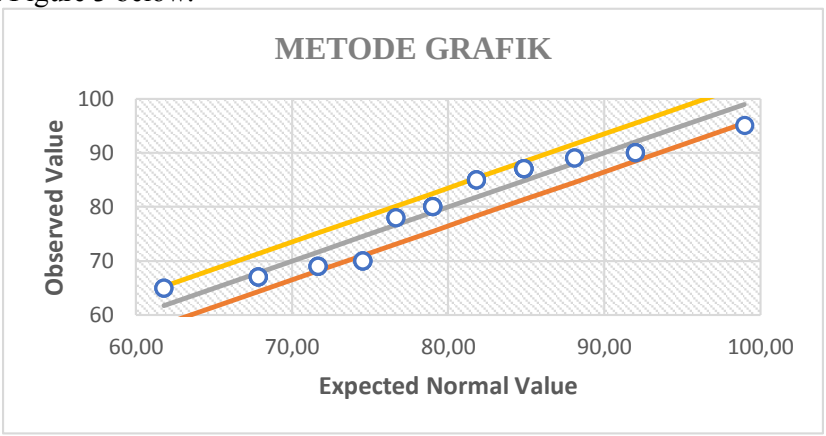

Figure 3. Q-Q plot with the horizontal axis is the expectation value and the vertical axis, namely observation data

ICSCCD 2018 p-ISSN 2716-1919 e-ISSN 2716-2907
Volume 1, January 2020 doi:10.31327/icusn-adri.v1i0.1152

4. Conclusion

From the explanation and example above can be drawn conclusions is Using this method the normal detection of data can be seen only through a graph and the results are more accurate than the Kolmogorov-Smirnov method is some cases. The implication of this research is that by using normal detection methods through graphs, it is expected to be implemented to test the normality of the data on experimental studies in coastal community.

References

[1] Ahmad F and Sherwani R A K 2015 *Pakistan J. Stat. Oper. Res.* **11** 331
[2] Boedec K L 2016 *J. Vet. Clin. Pathol.* **45** 648
[3] Yap B W and Sim C H 2011 *J. Stat. Comput. Simul.* **81** 2141
[4] Yazici B and Yolacan S 2007 *J. Stat. Comput. Simul.* **77** 175
[5] Fallo J O, Setiawan A and Susanto B 2013 *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika* (Yogyakarta: FMIPA UNY) p 978
[6] Razali N M and Wah Y B 2011 *J. Stat. Model. Anal.*. **2** 21
[7] Noiman S A, Brown L D, Buja A, Rolke W and Stine R A 2013 *J. Am. Stat.* **67** 249
[8] Nasrum A 2018 *Uji Normalitas Data untuk Penelitian* (Bali: Jayapangus Press)
[9] Johnson R A and Wichern W D 2002 *Applied Multivariate Statistical Analysis* 5th Ed (New Jersey: Prentice-Hall)