Is My Data Normal? Using Technology To Test For Normality

James Graziose
Palm Beach State College
3000 Saint Lucie Ave
Boca Raton, FL 33431
graziosj@palmbeachstate.edu

Abstract: In elementary statistics classes, many of our statistical tests that we perform on small data sets ($n < 30$) require the population from which the sample data was obtained be normally distributed. We explain to our students, a random variable X is normally distributed, or approximately normal, if the graph of the histogram is symmetric and bell-shaped or a normal-quantile plot which is linear. But in reality, we never obtain a perfect symmetric histogram or a normal quantile plot which is linear. Given a data set, which was obtained from a simple random sample whose distribution is unknown, we will apply two methods; a normal-quantile plot and Lilliefors test to assess the normality of the random sample.

1. Introduction

In elementary statistic classes, many of our statistical tests that we perform on small data sets $n < 30$, require that the population from which the sample data was obtained to be normally distributed. We explain to our students that a random variable X is normally distributed, or approximately normal, if: (1) The graph of the histogram is symmetric and bell-shaped or (2) a normal-quantile plot which is linear. Though in reality, we never obtain a perfect symmetric histogram or a normal-quantile plot which is linear.

Suppose that we have a data set which was obtained from a simple random sample from a population whose distribution is unknown. If this data set is relatively small, the histogram obtained from this data does not accurately represent the shape of the population. Other methods should be used to determine the normality which we now give.

2. The Normal-Quantile Plot

We briefly give the steps for constructing the normal-quantile plot.

Step 1: Arrange the data in ascending order.

Step 2: Compute the plotting position, Blom (1958)

$$f_i = \frac{i - 0.375}{n + 0.25}$$

where *i* is the index (the *i*th number in the list) and n is the number of observations. This value represents the expected proportion of observations less than or equal to the *i*th data value.

Step 3: Find the z-score corresponding to f_i : z-score = invnorm (f_i)

Step 4: Plot the observed values on the horizontal axis and the corresponding expected z-scores on the vertical axis.

Where:

$E =$ the expected of x_i (z-score) and is defined as:

$E\ x_i \sim \Phi^{-1}(f_i)$, where f_i is the expected proportions of observations $\leq x_i$.

And

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-x^2}{2}} dx \text{ , where } \mu = 0 \text{ and } \sigma = 1.$$

Example 1: The data in table 1 below represents the three year return of 19 randomly selected small-capitalization growth mutual funds. Is there evidence to support the belief that the variable "three-year rate of return" is normally distributed?

Table 1
Three year rate of return

Index *i*	Observed Value	f_i	Expected Z-score
1	15.8	0.0325	-1.85
2	16.7	0.0844	-1.38
3	18.2	0.1364	-1.10
4	18.4	0.1883	-0.88
5	18.4	0.2403	-0.71
6	18.5	0.2922	-0.55
7	19.2	0.3442	-0.40
8	19.5	0.3961	-0.26
9	21.3	0.4481	-0.13
10	22.2	0.5	0
11	22.6	0.5519	0.13
12	23.7	0.6039	0.26
13	23.7	0.6558	0.40

14	25.5	0.7078	0.55
15	27.0	0.7597	0.71
16	27.4	0.8117	0.88
17	28.5	0.8636	1.10
18	29.1	0.9156	1.38
19	29.6	0.9675	1.85

The normal-quantile plot of the data in table 1 is displayed in figure 1.

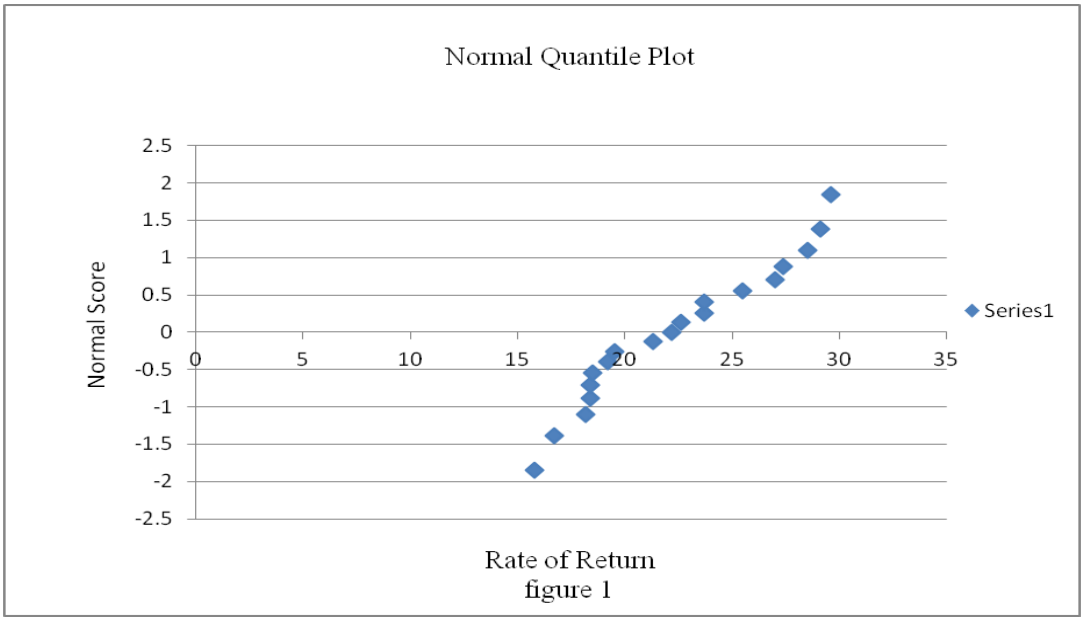

figure 1

As we can see, the normal-quantile plot in figure 1 does display curvature, and is roughly linear. The computed correlation between the observed value and the expected z-score is 0.974. Therefore, we can conclude that the three-year rate of return of small-capitalization growth mutual funds is approximately normally distributed.

3. Lilliefors Test for Normality

Let $X_1, X_2, \ldots, X_n$ be a random sample of data of size n associated with some unknown distribution function $F(X)$. The method Lilliefors (1967) used to determine the normality of a simple random sample is now outlined. For each value of X, we compute its corresponding z-score, Z_i, the sample cumulative distribution function $F_N(X)$, the cumulative normal distribution function $N^*(X)$, with mean $\mu = \overline{X}$, and variance $\sigma^2 = s^2$. The test statistic can now be determined using the following:

$T = \max | N^*(X) - F_N(X) |$. Now, if the value of the test statistic T is greater than the critical value found in table 3, then we can reject the null hypothesis H_0, and determine that our random sample of data is not from a normal distribution. This method is now illustrated in the following example.

Example 2: Sample data containing different ozone readings over the years have been compiled at the NASA website, **jwocky.gsfc.nasa.gov/index.html.** The ozone readings are given in Dobson units and the Nimbus satellite was selected. These readings were taken on each November 1st from 1978 through 1992, yielding the following fifteen sample readings: 329, 256, 212, 233, 185, 209, 243, 202, 203, 147, 360, 193, 148, 178, and 164.

The computed values from the data set of example 2 are shown in table 2.

Table 2
Ozone Readings in Dobson Units

Ozone Reading, X	Z_i	F(X)	$N^*(Z_i)$	$\| F_N(X) - N^*(X) \|$
147	-1.16	.066	.123	.057
148	-1.14	.133	.127	.006
164	-0.88	.200	.189	.011
178	-0.65	.266	.258	.008
185	-0.54	.333	.295	.038
193	-0.40	.400	.345	.055
202	-0.25	.466	.401	.065
203	-0.24	.533	.405	.128
209	-0.14	.600	.444	.156
212	-0.09	.666	.464	.202
233	0.26	.733	.603	.130
243	0.42	.800	.663	.167
256	0.63	.866	.736	.130
329	1.83	.933	.966	.033
360	2.34	1.000	.990	.010

From table 2, we can determine the test statistic, T = .202. Using table 3 of critical values calculated by Lilliefors (1967), with n =15 and $\alpha = 0.05$, we obtain a critical value of .220. Since our test statistic, T < .220, we do not reject H_0 therefore we can conclude that our data set is from a normal distribution.

4. Graphical Representation

With the aid of technology, for example using a graphing calculator, one can easily graph the sample cumulative distribution function and the cumulative normal

distribution function. Figure 2 illustrates Lilliefors method graphically of the ozone data from example 2. We can see from the graph that the sample cumulative distribution function (step function) of the sample data falls within the confidence interval bounds. Therefore, we can conclude, with 95 percent confidence, that the sample data follow a normal distribution.

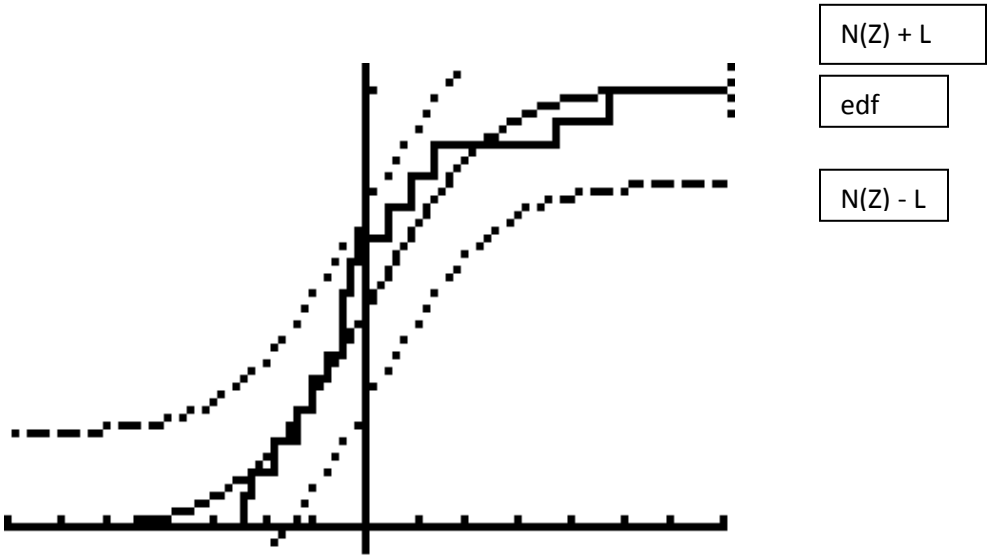

Figure 2
Lilliefors Graph of Ozone Data

Note: Where the edf (step function) represents the empirical distribution function.

The values of T given in the table 3, calculated by Lilliefors (1967) are critical values associated with selected values of N. Any value T which is greater than or equal to the tabulated value is significant at the indicated level of significance. These values were obtained as a result of Monte Carlo calculations, using 1,000 or more samples for each value of N.

Table 3
Critical values for
Level of significance for $T = \max | N^*(X) - F_N(X) |$

Sample Size N	.10	.05	.01
4	.352	.381	.417
5	.315	.337	.405
6	.294	.319	.364
7	.276	.300	.348
8	.261	.285	.331
9	.249	.271	.311
10	.239	.258	.294
11	.230	.249	.284
12	.223	.242	.275
13	.214	.234	.268
14	.207	.227	.261
15	.201	.220	.257
16	.195	.213	.250
17	.189	.206	.245
18	.184	.200	.239
19	.179	.195	.235
20	.174	.190	.231
25	.165	.180	.203
30	.144	.161	.187

Lilliefors (1967)

References

[1] Blom, G. (1958), *Statistical Estimates and Transformed Beta Variables*, New York: John Wiley.

[2] Calzado, M.E., and Scariano, S.M. (1999) "What is Normal, Anyway?" *The Mathematics Teacher*, 92, 682-689.

[3] Lilliefors, H.W. (1967), "On the Kolmogorov-Smirnov Test for Normality With Mean and Variance Unknown," *Journal of the American Statistical Association*, 62, 399-404.

[4] Looney, S.W., and Gulledge, T.R. (1985) "Use of the Correlational Coefficient With Normal Probability Plots," *The American Statistician*, 39, 75-79.

[5] Molin, P., and Abdi H. (1998) "New Tables and numerical approximations for the Kolmogorov-Smirnov/Lilliefors/Van Soest test of normality." Technical report, University of Bourgogne.

[6] Sullivan, M. (2004), *Statistics Informed Decisions Using Data*, New York: Prentice Hall.