How Normal Is Normal?

Using a Q-Q Plot

BY ALEX T.C. LAU

Q: How can I determine if my data comes from a normal distribution?

A. A quantile-quantile, or Q-Q, plot can be used to examine if a data set is approximately normal.

A lion's share of statistics interpretation and associated decision making are based on the assumption that the universe from which the limited data set is obtained, or the statistics calculated from the data set, can be adequately represented (modeled) by the Gaussian, which is more commonly known as the normal distribution. There is a plethora of techniques that can be used to validate the reasonableness of this normal assumption. Most techniques will require a commercial statistical software package to carry out the necessary computations and plots. This article describes a graphic technique that can be used to visually determine if the data are approximately normally distributed. The technical name for this technique is the Q-Q plot.

The Q-Q plot is a graphical method for studying how well the underlying distribution from which the dataset is collected can be approximated by the normal model. It is equivalent to the classical normal probability plot but, unlike the latter, no specialized scale or probability paper is required. This plot can be easily implemented in a spreadsheet tool such as Excel using the NORMSINV function. The data can be deemed to be "adequately" normal if most of the points in the plot lie roughly along a straight line. In addition to judgment of data normality, other salient features associated with the Q-Q plots are:

- The y-axis is in the original units of the data,
- Potential outlier(s) can be visually identified as the point(s) that deviate significantly from the approximate straight line along which most of the data lie,
- The y-intercept of the approximate straight line is the median of the data set, and
- The slope of the approximate straight line is an indication of the magnitude of the data set standard deviation, where a steep slope represents a large standard deviation and a shallow slope represents a small standard deviation.

A simple description of how to construct a Q-Q plot is outlined below. The Q-Q plot procedure is as follows:

1. Order the data from smallest to largest (n = total number of observations).

2. Create an index i next to the ordered data where i will take on values from 1 through n, with the lowest value assigned $i = 1$ and the highest assigned $i = n$.

3. Calculate $f_i = (i - 0.5)/n$ for each observation. This is a rank plotting position for the Q-Q plot.

4. Obtain from the cumulative distribution version of a standard normal distribution table ($\mu = 0$, $\sigma = 1$) the value of z_i for each f_i. An easier approach is to use the Excel spreadsheet function NORMSINV function to compute the z_i values as shown in Table 1. Pair it to the observation with index i for plotting later.

5. Plot each observation value on the y-axis against its z_i value obtained in step 4 on the x-axis using ordinary linear graph paper. This creates the Q-Q plot (see Figure 1).

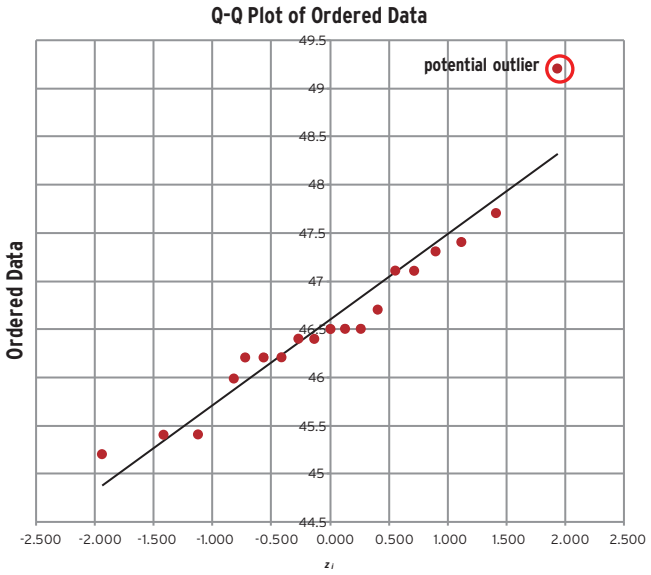
The next step is to visually examine the plot for approximate linearity. If the Q-Q plot pattern is linear, or nearly so, the data distribution is well approximated by the normal model. Significant deviation from linearity should serve as a signal for potential failure of the normality assumption.

Interested readers are referred to ASTM D6299, Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance, for a detailed description of the Q-Q plot as well as how to calculate an associated A-D (Anderson-Darling) statistic to assess data normality.

Table 1 – Data for Q-Q Plot

Original Data	Ordered Data	Index *i*	$f_i = (i-0.5)/n$	z_i =NORMSIN(f_i)
46.4	45.2	1	0.026	-1.938
46.5	45.4	2	0.079	-1.412
45.4	45.5	3	0.132	-1.119
46.4	45.9	4	0.184	-0.899
46.7	46.2	5	0.237	-0.716
47.1	46.2	6	0.289	-0.555
45.2	46.2	7	0.342	-0.407
45.5	46.4	8	0.395	-0.267
46.2	46.4	9	0.447	-0.132
47.1	46.5	10	0.500	0.000
47.4	46.5	11	0.553	0.132
45.9	46.5	12	0.605	0.267
46.2	46.7	13	0.658	0.407
46.2	47.1	14	0.711	0.555
47.3	47.1	15	0.763	0.716
46.5	47.3	16	0.816	0.899
49.2	47.4	17	0.868	1.119
46.5	47.7	18	0.921	1.412
47.7	49.2	19	0.974	1.938

Figure 1 – Q-Q Plot of Ordered Data* versus Z_i**



*Ordered Data is second column from left in Table 1.
** Z_i is fifth (last) column from left in Table 1.

ALEX T.C. LAU, *TCL Consulting, Whitby, Ontario, Canada, is chairman of Subcommittees D02.94 on Quality Assurance and Statistics and D02.01.0B on Precision,, which are part of ASTM Committee D02 on Petroleum Products and Lubricants. An ASTM International fellow, Lau is also a member of Committees E11 on Quality and Statistics, E36 on Accreditation and Certification, and F08 on Sports Equipment and Facilities.*

Dean V. Neubauer, Corning Inc., Corning, N.Y., is an ASTM International fellow, chairman of E11.90.03 on Publications and coordinator of the DataPoints column; he is immediate past chairman of Committee E11 on Quality and Statistics.

snonline

Get more tips for ASTM standards development at www.astm.org/sn-tips.

Find other DataPoints articles at www.astm.org/standardization-news/datapoints.