A Graphical Tool for Assessing Normality

Martin L. HAZELTON

Interpretation of normal probability plots is not always straightforward for the inexperienced data analyst. In the finance literature a plot of empirical and fitted normal densities on the log scale is frequently preferred as a graphical diagnostic for normality. This article describes the construction of this type of plot, and suggests a refinement that can facilitate its interpretability with small samples. A Monte Carlo test for normality arises naturally as a by-product of this methodology.

KEY WORDS: Anderson–Darling test; Cramer–von-Mises test; Kernel density estimation; Monte Carlo test; Shapiro–Wilk test.

1. INTRODUCTION

The normal probability plot is a commonly used tool for informal assessment of the normality of distributions of observed data or residuals. In such a graph the ordered sample is plotted against corresponding normal scores. An approximate straight line relationship is consistent with underlying normality while curvature in the plot provides evidence to the contrary. Although the idea is straightforward, correct interpretation of normal probability plots requires some care and attention. For the inexperienced data analyst it can prove difficult to judge whether a plot is sufficiently nonlinear to cast doubt on the normality of the sample. It requires a further level of expertise to glean the form of nonnormality (e.g., skewness, heavy tails) from this type of plot. It can be argued that use of the quantile scale in normal probability plots makes them less intuitive than fitted line or residual plots.

Modeling stock returns in finance is a problem where assessment of data normality is important. The celebrated Black and Scholes (1973) model assumes that the distribution of log returns is normal, but real financial data can display a variety of nonnormal features depending on the sampling frequency. See Rydberg (2000), for example. Graphs illustrating nonnormal returns data are not uncommon in the finance literature, but the use of normal probability plots appears rare. Authors in this area generally prefer to plot an "empirical density" for the data, against which a fitted normal density curve is superimposed for comparison. A log-scale is used on the ordinate axis so as to clearly exhibit the tails of the distribution of the data. Such log-density plots can be effective graphical diagnostics for normality, but they can be misleading when the sample size is modest unless the empirical and fitted normal densities are carefully defined.

The purpose of this article is to bring the advantages of log-density diagnostic plots to the attention of statisticians outside the field of finance, and to describe how the construction of such plots may be refined in order to facilitate their correct interpretation when the sample size is small. The next section describes the plotting methodology in detail, paying particular emphasis to the manner in which the empirical density is defined. We then demonstrate the problems that arise when employing this type of plot with small samples, and describe the aforementioned refinement to the plot. Section 3 develops a Monte Carlo test for normality which arises as a natural by-product of the graphical tool. Through experiments with real and simulated data we show that this test is more sensitive than standard alternatives (such as the Shapiro–Wilk, Anderson–Darling, or Cramer–von-Mises tests) at detecting nonnormality in the tails of the distribution of the data.

Martin L. Hazelton is a Senior Lecturer in Statistics, School of Mathematics and Statistics M019, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia (E-mail martin@maths.uwa.edu.au). The author thanks Tony Pakes for introducing him to relevant finance literature, and is grateful to an associate editor and anonymous referee for their helpful comments.

2. THE BASIC PLOT AND ITS REFINEMENT

We wish to assess the normality of a random sample $X_1, \ldots, X_n$. The basic plotting tool of interest is constructed as follows. An "empirical density" is computed from the data and plotted as points on a graph using a log scale on the ordinate axis, as described in the following. The logarithm of a normal density, with mean and variance estimated from the data, is plotted as a line for comparison. We illustrate this methodology using daily log returns for IBM stock in 1980 and 1981. As with many other types of assets, the log returns show very little serial correlation and so it is not unreasonable to regard these data as a random sample for the purposes of producing the graphical diagnostic; see Rydberg (2000). The plot for the IBM data is given in Figure 1. This graph has much of the appeal of a fitted line plot, allowing easy assessment of major departures of the data from normality. In the case at hand, the distribution of the data clearly has tails

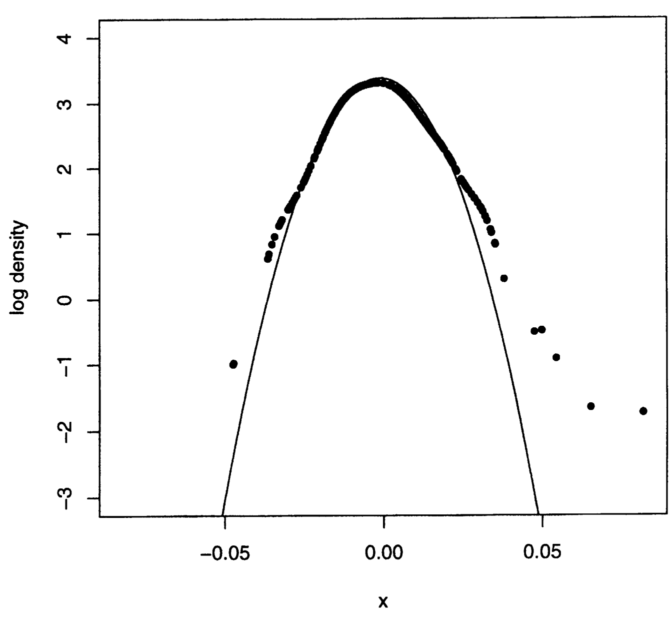

Figure 1. Plot of the empirical density of daily log returns for IBM stock in 1980–1981 (points), and fitted normal density (line). The ordinate axis is on a log scale.

 DOI: 10.1198/0003130032341

that are too fat to suggest normality, a typical observation when dealing with log returns.

The calculation of the "empirical density" of the data is a fundamental element in constructing the log-density plots. We use the phrase "empirical density" because it is widely employed in the finance literature when describing these plots. In that context it can refer to a number of different nonparametric methods for estimating a density (but not to the probability mass function that places probability mass $1/n$ at each datum). In practice the empirical density is usually calculated using a histogram (scaled to have unit area) or its generalization, a kernel density estimate. The plots constructed using kernel density estimation are the more appealing in the author's experience, and so we concentrate on this version of the methodology.

The kernel density estimator computed from the data $X_1, \ldots, X_n$ is defined by

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \qquad (1)$$

where $K_h(x) = h^{-1}K(x/h)$. Here K is a kernel function which is a symmetric probability density function such as the normal density, and h is the smoothing bandwidth. If K is the normal density then one can think of $\hat{f}$ as being built from n bell shaped "blocks", each centered at a data point and carrying probability mass $1/n$. If the kernel function is rectangular and each X_i is replaced by the nearest bin midpoint, then we obtain the usual histogram. A good account of kernel density estimators (and their advantages over histograms) is provided by Silverman (1986) or Simonoff (1996). The empirical density in the log density plots is defined as the kernel density estimate evaluated at the data, so that the points plotted in Figure 1 and the like are $(X_1, \log[\hat{f}(X_1)]), \ldots, (X_n, \log[\hat{f}(X_n)])$. It should be noted that the smoothing process implicit in (1) implies that $\hat{f}(X_i)$ and $\hat{f}(X_j)$ are highly dependent if X_i and X_j are close. As a result, the "residuals" in plots like Figure 1 are not independent, and so runs of points a little way above or below the fitted log density are not in themselves evidence of a lack of normality in the data.

To compute the empirical density the bandwidth h must be chosen. Bandwidth selection for kernel density estimation has been the subject of a great deal of research and a number of sophisticated methods have been developed. See Wand and Jones (1995), for instance. Although one of these advanced bandwidth selectors could be applied in the current context, we advocate use of the simple *normal scale bandwidth*. This bandwidth is optimal (in the sense that it minimizes asymptotic integrated mean squared error) if the underlying distribution of the data is normal. If K is the standard normal density (as we shall assume henceforth), then the normal scale bandwidth is given by $h_{\text{NS}} = 1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ is an estimate of the standard deviation. See Silverman (1986). We regard h_{NS} as adequate because it is close to optimal in challenging situations where the distribution of the data is not obviously nonnormal, while the choice of bandwidth is not crucial in other cases.

Assessment of normality from the plotted empirical density is facilitated by comparison with a fitted normal density curve (on the log scale). The normal density can be fitted by estimating mean, μ, and variance, σ^2, by the respective sample moments, but the presence of outliers can inflate the sample variance (in particular) and thus have an extreme effect on the fitted density. We therefore advocate that robust estimates of location and spread be used, as appears to be common practice for production of the log-density plots in the finance literature. Our preference is to use the 20% trimmed mean as an estimate of μ, and mean absolute deviation, multiplied by the factor 1.4826 (to ensure asymptotically normal consistency) as an estimate of σ. The log-density plot produced using the empirical density, defined above, and robust fitted normal curve is what we henceforth refer to as the "basic plot."

We now propose a refinement to the basic plot that modifies the fitted normal curve. There is a bias inherent in kernel density estimation that has the effect of inflating the spread of the empirical density in comparison to the underlying density, f. The magnitude of this inflation is dependent upon the bandwidth. More specifically, suppose that f is normal with mean μ and variance σ^2. Then (assuming that K is a standard normal density)

$$\begin{aligned} \mathrm{E}[\hat{f}(x)] &= \int K_h(x-y)\phi_\sigma(y-\mu)\,dy \\ &= K_h * \phi_\sigma(x-\mu) \\ &= \phi_{\sqrt{\sigma^2+h^2}}(x-\mu), \end{aligned}$$

where $\phi_\sigma(\cdot)$ is the density function for a normal distribution with zero mean and variance σ^2, and $*$ denotes a convolution. It follows that if f is normal, then the empirical density estimates a $N(\mu, \sigma^2 + h^2)$ density rather than a $N(\mu, \sigma^2)$ density. Hence the most appropriate fitted normal distribution for comparison with the empirical density should have variance $\hat{\sigma}^2 + h^2$ rather than $\hat{\sigma}^2$, where a hat indicates a robust estimate of the corresponding population quantity. If the sample size is large, then our bandwidth of choice, $h_{\text{NS}} = 1.06\hat{\sigma}n^{-1/5}$, will be small and there will be little loss in ignoring the inflation in spread. For instance, when $n = 500$ the inflation in standard deviation from

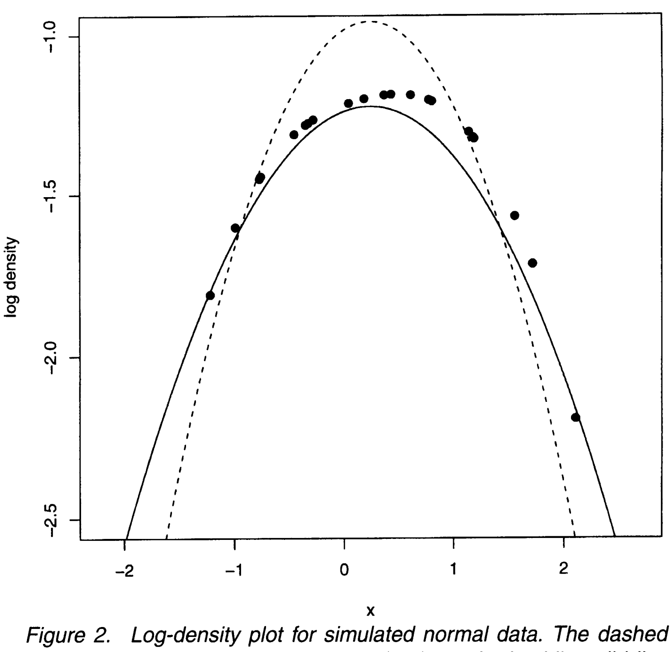

Figure 2. Log-density plot for simulated normal data. The dashed line is the fitted normal curve using the basic method, while solid line corresponds to the refined methodology.

Table 1. Power of Normality Tests for Cauchy Data. Simulation standard errors are given in parentheses.

n	MCLD	SW	AD	CvM
20	0.897 (0.006)	0.859 (0.007)	0.860 (0.007)	0.853 (0.007)
50	0.997 (0.001)	0.997 (0.001)	0.997 (0.001)	0.997 (0.001)
200	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

$\hat{\sigma}$ to $(\hat{\sigma}^2 + h^2)^{1/2}$ is about 4.6%, while the inflation is down to 1% when $n = 5{,}000$. However, for small sample sizes the fitted normal curve should account for the bias in the empirical density.

In light of this discussion, we advocate using a normal log-density fitted by robust mean estimate $\hat{\mu}$ and variance $\hat{\sigma}^2 + h^2$ in the log-density plot. To illustrate the importance of this refinement for modest sample size we generated a random sample of 25 observations from a standard normal density, from which we computed the empirical density using bandwidth $h_{\text{NS}} = 0.67$. The empirical density is plotted (on the log scale) in Figure 2 along with fitted normal curves using the basic method (the dashed line) and refined method (solid line) implemented using robust estimates $\hat{\mu} = 0.26$ and $\hat{\sigma} = 1.20$. The fit of the dashed line is quite poor and suggests that the data might be drawn from a distribution with fatter tails than the normal, despite the fact that a Shapiro–Wilk test (Shapiro and Wilk 1965) indicates no reason to doubt normality (p value $= 0.85$). On the other hand, the solid line indicates the normality of the data. Note that the run of data points slightly above the refined curve is not indicative of a lack of fit because of the dependency in the empirical density points, as discussed earlier.

3. A MONTE CARLO TEST

The ideas behind the plot described above can be adapted in a simple fashion to provide a test for normality. To develop a test statistic, we define the ith residual in the log-density plot to be

$$e_i = \log(\hat{f}(X_i)) - \log(\hat{\phi}(X_i)) \qquad (i = 1, \ldots, n),$$

where $\log(\hat{f}(X_i))$ is the empirical density for X_i, and

$$\hat{\phi}(X_i) = \phi_{\sqrt{\hat{\sigma}^2 + h^2}}(X_i - \hat{\mu})$$

is the corresponding fitted value. The sum of squared residuals, $S = \sum_{i=1}^n e_i^2$, is then a natural measure of the overall discrepancy between the empirical and fitted densities. Hence S is an appropriate statistic for testing the null hypothesis that the data are drawn from a normal distribution.

We can compute a p value for an observed test statistic, s_{obs}, using a standard type of Monte Carlo procedure. The basic idea behind this approach is to compare s_{obs} with values of the test statistic computed from datasets of size n simulated under the null hypothesis. If s_{obs} is large in comparison to these simulated test statistics, then doubt is cast on the normality of the observed data. More specifically, if we simulate N datasets from a suitable normal distribution and compute the test statistic for each one, then the Monte Carlo p value is the proportion of those test statistics that are larger than s_{obs}. This process is described in algorithmic format below.

Table 2. Power of Normality Tests for Data from the t Distribution with Six Degrees of Freedom. Simulation standard errors are given in parentheses.

n	MCLD	SW	AD	CvM
20	0.161 (0.007)	0.159 (0.007)	0.114 (0.006)	0.100 (0.006)
50	0.319 (0.009)	0.285 (0.009)	0.207 (0.008)	0.203 (0.008)
200	0.776 (0.008)	0.687 (0.009)	0.540 (0.010)	0.531 (0.010)

Algorithm for Computing the Monte Carlo p value

1. Generate N independent samples, each of size n, from a normal distribution with zero mean and variance $\hat{\sigma}^2$. Let $X_1^{(j)}, \ldots, X_n^{(j)}$ denote the jth such sample.

2. Compute

$$e_i^{(j)} = \log(\hat{f}^{(j)}(X_i^{(j)})) - \log(\hat{\phi}^{(j)}(X_i^{(j)})) \qquad (i = 1, \ldots, n \quad j = 1, \ldots, N),$$

where $\hat{f}^{(j)}$ and $\hat{\phi}^{(j)}$ are the empirical and fitted normal densities obtained from the jth sample. Calculate the jth sample test statistic, $S^{(j)} = \sum_{i=1}^n (e_i^{(j)})^2$.

3. Calculate the Monte Carlo p value, $P = \#(S^{(j)} \geq s_{\text{obs}})/N$.

Note that the choice of zero mean for the normal distribution at Step 1 is arbitrary. The test statistic S is invariant under a change of location of the dataset.

The choice of N when implementing this algorithm is a balance between minimizing Monte Carlo (i.e., simulation induced) error, which requires large N, and computing time. The Monte Carlo standard error is $\sqrt{p(1-p)/N}$ when the p value is p. We regard $N = 1{,}000$ as generally sufficient, giving a standard error which is bounded above by 0.016 and evaluates to 0.007 when $p = 0.05$. A larger value of N is only likely to be necessary when the initial Monte Carlo p value is within two or three standard errors of the specified significance level. We used $N = 1{,}000$ when generating all the numerical results which follow.

In order to compare the power of our test with that of some common tests for normality, we performed a small simulation study. In this study we considered data from three distributions, namely a Cauchy distribution, a t distribution with six degrees of freedom, and an extreme value (Gumbel) distribution. We considered three sample sizes, $n = 20$, $n = 50$, and $n = 200$, and three alternative normality tests: the Shapiro–Wilk test (Shapiro and Wilk 1965), the Anderson–Darling test (Stephens 1974), and the Cramer–von-Mises test (e.g., D'Agostino and Stephens 1986). The power for each test was estimated for each com-

Table 3. Power of Normality Tests for Data From the Extreme Value (Gumbel) Distribution. Simulation standard errors are given in parentheses.

n	MCLD	SW	AD	CvM
20	0.207 (0.008)	0.324 (0.009)	0.243 (0.009)	0.219 (0.008)
50	0.416 (0.010)	0.698 (0.009)	0.558 (0.010)	0.540 (0.010)
200	0.913 (0.006)	1.000 (0.000)	0.992 (0.002)	0.991 (0.002)

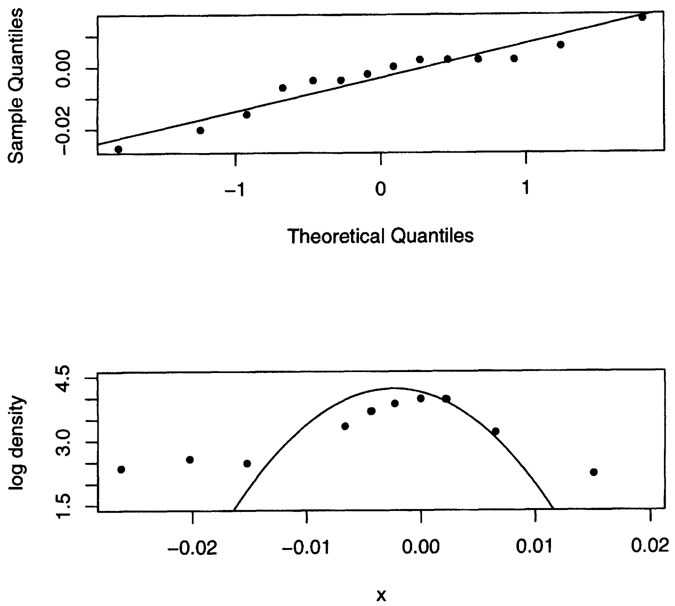

Figure 3. Normal probability plot (top panel) and log-density plot (bottom panel) for weekly log returns on AT&T common shares.

bination of distribution and sample size from 2,500 simulated datasets. The test size was set at 0.05 throughout. The results are displayed in Tables 1–3, where the tests are labeled MCLD for our Monte Carlo log-density test; SW for the Shapiro–Wilk test; AD for the Anderson–Darling test; and CvM for the Cramer–von-Mises test.

For the Cauchy and t_6 distributions, the power of MCLD is no lower than that of the alternative tests for any of the sample sizes considered, and is significantly higher in some cases. This demonstrates the sensitivity of MCLD to fat tailed departures from normality, a property that we might have expected given the use of the log-density scale. In the case of the extreme value distribution, where skewness is the most prominent nonnormal feature, the situation is reversed and the performance of MCLD suffers in comparison to the other tests.

We conclude with a real data example to illustrate the application of the log-density plot and the Monte Carlo test to a small dataset. The data are 14 weekly log returns on AT&T common shares during the period June 28 to October 1, 1979. A normal probability (Q-Q) plot and log-density plot for these data are given in Figure 3. The curvature in the former plot is somewhat suggestive of a lack of normality, particularly when viewed in comparison to a reference line. Nonetheless, for an inexperienced data analyst in particular, the log-density plot illustrates the fatter-than-normal tails of the data's distribution more clearly. Neither the Shapiro–Wilk test nor the Anderson–Darling test gives reason to doubt the normality of the data, providing p values of 0.32 and 0.14, respectively. However, the Monte Carlo test p value is 0.023 (with Monte Carlo standard error 0.005) giving clear evidence of a lack of normality, a finding that is in keeping with many other analyses of financial log return data (see, e.g., Rydberg 2000).

4. CONCLUSION

In the finance literature, log-density plots are a popular and intuitive alternative to normal probability plots. We have described the construction of a basic log-density plot, and have suggested a refinement whereby the plotted empirical density is compared with an appropriately scaled normal density on the log-scale. The refined log-density plot is particularly effective at displaying departures from normality in the tails of the data. A Monte Carlo test for normality arises as a natural by-product of this methodology. This test is especially sensitive to fat-tailed departures from normality, and performs better than well-known alternatives (such as Shapiro–Wilk, Anderson–Darling, and Cramer–von-Mises tests) in this regard. Code for implementing both the refined plot and the Monte Carlo test in the statistical software package R (Ihaka and Gentleman 1996) is available on the WWW at http://www.maths.uwa.edu.au/~martin/ldplot.htm.

[Received November 2002. Revised July 2003.]

REFERENCES

Black, F., and Scholes, M. (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637–654.

D'Agostino, R., and Stephens, M. (1986), *Goodness-of-Fit Techniques*, New York: Decker.

Ihaka, R., and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Rydberg, T. (2000), "Realistic Statistical Modelling of Financial Data," *International Statistical Review*, 68, 233–258.

Shapiro, S., and Wilk, M. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52, 591–611.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman & Hall.

Simonoff, J. (1996), *Smoothing Methods in Statistics*, New York: Springer.

Stephens, M. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730–737.

Wand, M., and Jones, M. (1995), *Kernel Smoothing*, London: Chapman & Hall.