Problem 1: Test for normal distribution and transformation

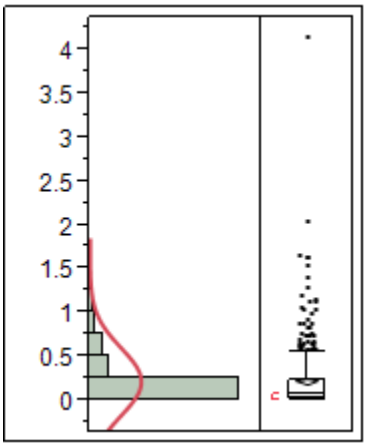
The first step in data analysis is often to test the data for conformance with a normal distribution. The distribution of the data (along with other characteristics of constant variance and independence of observations) determines the types of statistical tests that can be applied to the data. Use of parametric tests such as the Student's t-Test, analysis of variance, and linear regression requires that the data follow a normal distribution, that variance is the same for all groups (e.g., different stations or different periods), and that observations are independent of each other. Independence can be evaluated by examining autocorrelation (see section 7.3.6). Most statistics software offer diagnostics and alternative analyses for groups with different variance (e.g., Levene's and Brown-Forsythe which are both available in JMP). This example will demonstrate testing raw data and transformed data for conformance with a normal distribution.

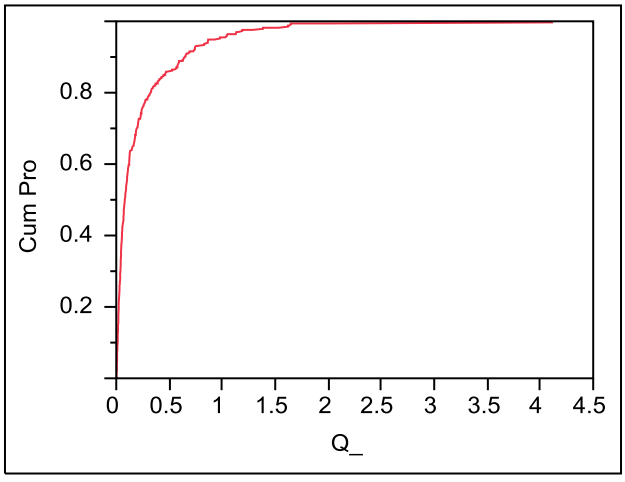
Using Dataset 1 in file Sampledata.xlsx, test flow data at Station 1 (Q_1) for conformance with a normal distribution and, if the data do not follow a normal distribution, test a $\log_{10}$ transformation. A $\log_{10}$ transformation is done by simply taking the base$_{10}$ logarithm of each value, which can be done as a function in a spreadsheet or by calculating a new variable in a statistics program. Note that if zero values are legitimate for a particular variable (e.g., flow or load), the $\log_{10}$ is undefined and will result in an error or a missing data value. In such cases, alternative transformations (e.g., log(value + 1) may be considered.

Results:

Step 1: test the raw data

Q_1





Summary Statistics

Mean	0.2256577
Std Dev	0.3840104
Std Err Mean	0.0210436
N	333
Skewness	4.5961049
Kurtosis	34.420844
CV	170.1739
Median	0.081

Fitted Normal
Goodness-of-Fit Test
Shapiro-Wilk W Test

W	Prob<W
0.561197	<.0001*

Note: Ho = The data are from the Normal distribution. Small p-values reject Ho.

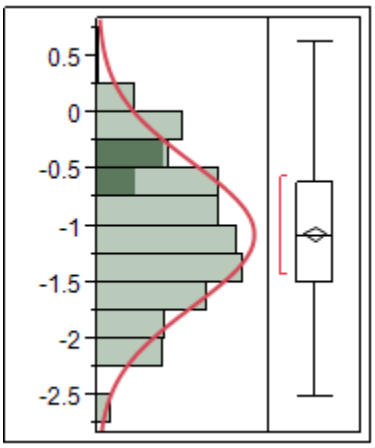
Fitted LogNormal
Goodness-of-Fit Test
Kolmogorov's D

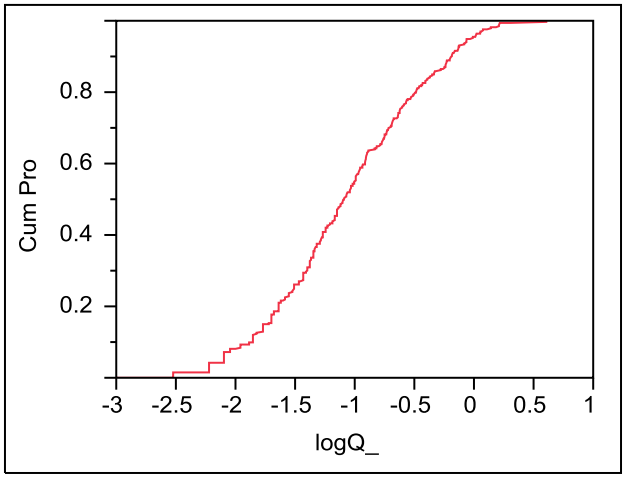
D		Prob>D
0.035622	>	0.1500

Note: Ho = The data are from the LogNormal distribution. Small p-values reject Ho.

The above graphs suggest that the data are not from a normal distribution. The histogram shows that flow data are highly right-skewed, with a preponderance of very low values and a few high extremes; this is confirmed by the cumulative frequency plot that shows about 80 percent of observations are $\leq$ 0.5 m^3/s. Descriptive statistics confirm this pattern. The mean (0.22 m^3/s) is much higher than the median (0.08 m^3/s); the skewness statistic of 4.96 shows that the data are strongly skewed. Furthermore, the Shapiro-Wilk W Test results in a P value $\leq$0.001, requiring the rejection of the hypothesis that the data are from a normal distribution. Finally, the Kolmogorov D statistic suggests that the hypothesis that the data are from a lognormal distribution cannot be rejected.

Step 2: test the $\log_{10}$ transformation
logQ_1





Summary Statistics

Mean	-1.070498
Std Dev	0.6334468
Std Err Mean	0.0347127
N	333
Skewness	0.0158462
Kurtosis	-0.552514
CV	-59.17309
Median	-1.091515

Fitted Normal
Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.991468	0.0517

Note: Ho = The data are from the Normal distribution. Small p-values reject Ho.

The log10 transformation appears to have yielded a dataset that more closely conforms to a normal distribution, as shown by both the histogram and the probability plot. The mean (-1.07) is much closer to the median (-1.09) and the Shapiro-Wilk W statistic does not reject the hypothesis that the log-transformed data are from a normal distribution.