ABSTRACT
Many statistical tests require data to be approximately normally distributed. Usually, the first step of data analysis is to test the normality. Also, we often test the normality of residuals after fitting a linear model to the data in order to ensure the normality assumption of the model is satisfied. SAS has offered four statistical tests that provide an easy way to test the normality. However, we should be cautious when we use these tests due to their limitations. In some cases, we may draw incorrect conclusions by only looking at the test statistics and p-values. Graphical methods are powerful in displaying distribution characteristics of the data and can serve as a useful tool in checking the normality. Combining graphic methods and statistical tests will improve our judgments on the normality of the data. In this paper, I will present these methods SAS uses by applying them to the real data from a clinical trial.

INTRODUCTION
The first step of data analysis usually involves making distributional assumption about the data. If the data is considered truly a sample from some classes of probability distributions, we cannot only summarize the data compactly based the approximate distribution, but also carry out proper statistical procedures to gain valuable inferences. Furthermore, understanding the data distribution can sometimes shed light on the underlying mechanisms for generating the data (Chambers and Cleveland, etc. 1983). If the specified distributional assumption about the data is not valid, the analyses based on those assumptions will be invalid and sometimes lead to the incorrect conclusions. Many powerful statistical methods require approximate normality about the data, i.e. the data is a sample from a normal distribution. Therefore, checking the normality assumption is critical in the process of data analysis.

If we look at the general linear models (GLM), the dependent variables do not necessarily have normal distribution. But the error term in the model should have a standard normal distribution if the functional part of the model is specified correctly. Since residuals after model fitting approximate the actual errors, we should check the normality of the residuals to assure the validity of statistical tests using the GLM.

SAS has implemented four commonly used normality tests in PROC UNIVARIATE and PROC CAPABILITY. These tests are generally powerful (sensitive) to detect abnormality in the data. It is easy to automate the analysis process based on test p-values. SAS also has offered several graphic methods in the same procedures. Graphical methods such as Q-Q plot and density plot are powerful tools to check the distributional assumption. In many cases, we can see the detail position of every data point as well as gain overall perception about the data using graphical methods, while normality tests are easily influenced by the presence of a few high-leverage points or outliers and plateaus, segments of the data.
In this paper, I will review the four normality tests and two graphical methods, and illustrate the advantage and disadvantage of both methods by performing the normality test on a set of real clinical trial data.

NORMALITY TESTS USED IN SAS
Shapiro-Wilk test checks the normal assumption by constructing W statistic, which is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance (Shapiro and Wilk, 1965). To perform the test, the W statistic is constructed by considering the regression of ordered sample values on corresponding expected normal order statistics, which is linear for a sample from a normally distributed population (Royston, 1992). W is positive and less than or equal to one. Small values of W lead to the rejection of normality, while being close to 1 indicate normality of the data. However the distribution of W is highly skewed. The large values of W (W=0.9) may be considered small and lead to the rejection of normality (SAS Institute, 1999). So, the calculation of W has the strong impact on the test result. Against the same set of data, the value of W from Shapiro-Wilk test in SAS version 8.02 will be slightly different from that in SAS version 6.12 because the modification of algorithm for calculating W. Sometimes this slight difference may lead to the rejection of normality that has been established before. The original W statistics is valid for the sample sizes between 3 and 50, but Royston extended the test to the sample size of 2000 by transforming the null distribution of W to approximate normality. Unlike some other normality tests, Shapiro-Wilk test does not require specifying the mean and variance in advance and it is very powerful to detect the small departure from normality. But it will not indicate the source of abnormality.

Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (EDF). The test statistic, K-S D, is the largest vertical distance between the distribution function (F(x)) and the EDF (Fn(x)), which is a step function that takes a step of height 1/n at each observation. To test normality, the K-S D statistic is computed using the sample data against a normal distribution with mean and variance equal to the sample mean and variance (SAS Institute, 1999). The attractive feature of this test is that the distribution of the K-S D does not depend on the underlying distribution function being tested, therefore it is considered as non-parametric and distribution free. However, it is not
reflecting that the residuals are asymmetric and skewed to the right. The slope on the left part is very flat, which is
Q-Q shows the overall departure from the straight line (Figure 1). The slope of the curve increases from left to right,
density and the very small p-value of Shapiro-Wilk test. We can further investigate the normality using normal Q-Q plot and
tests are also done at the same time in SAS 8.2. The normality is strongly rejected as we see the small value of W
After fitting ANOVA model where the dependent variable is the percent change in IEF and independent variables are
(1) Percent change in IEF
Normality of residuals
Normal quantile-quantile plot (Q-Q plot) is the most commonly used and effective diagnostic tool for checking
normality of the data. It is constructed by plotting the empirical quantiles of the data against corresponding quantiles
of the normal distribution. If the empirical distribution of the data is approximately normal, the quantiles of the data
will closely match the normal quantiles, and the points on the plot will fall near the line y=x. It is impossible to fit a
straight line in Q-Q plot for the real data due to the fact that the random fluctuations will cause the points to drift away
and aberrant observations often contaminate the samples. Only large or systematic departures from the line indicate
the abnormality of the data. The points will remain reasonably close to the line if there is just natural variability.
Therefore, the straightness of the normal Q-Q plot helps us to judge whether the data has the same distribution
shape as a normal distribution, while shifts and tilts away from the line y=x indicate differences in location and
spread, respectively.
Another graphical method for normality test is the kernel density plot that portrays the distribution of data directly. In
order to get the plot, we first have to perform statistical density estimation, which involves approximating a
hypothesized probability density function from the observed data. Kernel density estimation is a nonparametric
technique for density estimation in which a known density function (kernel) is averaged across the observed data
points to create a smooth approximation. After plotting the density function, we can easily check the normality by
comparing the shape of resulting plot with the bell-shaped curve of normal distribution. Selection of kernel function
and bandwidth determine the smoothness of the plot, which sometimes makes the plot look different and in turn
affects our judgment. In SAS, PROC KDE uses a normal density as the kernel, and its assumed variance determines
the smoothness of the resulting estimate (SAS institute, 1998).
CASE STUDY
The data to be analyzed is from a double-blinded, randomized, placebo-controlled trial of a pharmacotherapy for
incontinence (n=458). Its primary objective is to compare the treatment group with the placebo group in two kinds of
measurements: percent change in Incontinence Episode Frequency (IEF) and the change in patient’s quality of life
score from the baseline to the endpoint using a validated condition specific quality of life questionnaire (Incontinence
Quality of Life, IQOL). The analysis plan proposed to fit the data using a linear model first, then check the validity of
normality assumption. If normality assumption does not hold, it was proposed to use a nonparametric method
instead. It is important to make correct judgment on normality to optimize the statistical power.
Normality of residuals
(1) Percent change in IEF
After fitting ANOVA model where the dependent variable is the percent change in IEF and independent variables are
therapy group, investigator and baseline stratification. The normality of residuals is first checked using normality
tests (Table 1). Since the sample size is less than 2000, Shapiro-Wilk test is the choice even though three other
tests are also done at the same time in SAS 8.2. The normality is strongly rejected as we see the small value of W
and the very small p-value of Shapiro-Wilk test. We can further investigate the normality using normal Q-Q plot and
density function curve. Although a relatively large portion of the data follows a straight line in the middle section, the
Q-Q shows the overall departure from the straight line (Figure 1). The slope of the curve increases from left to right,
reflecting that the residuals are asymmetric and skewed to the right. The slope on the left part is very flat, which is
caused by a lot of ties in the data. The empirical density curve (Figure 2) re-assures the asymmetric distribution. The odd appearance at the peak of the curves shows the two clusters of data points with the high densities. Obviously, the ANOVA model cannot be used. An alternative nonparametric test should be used to analyze IEF.

(2) Change in IQOL
Similarly, test statistics for normality of the residuals from an ANCOVA model where change in IQOL scores are the dependent variables and are therapy group, investigator baseline IQOL scores are independent variables are presented in Table 2. The Shapiro-Wilk test rejects the normality. Although the W statistic (W=0.9879) is close to 1, the p-value is small due to the skewness of W. We can investigate the issue by looking at the normal Q-Q plot of the residuals (Figure 3), the majority of the data points follow the straight line, while at the upper right section, the data points departure from the line y=x. This small deviation leads to rejection of the normality by Shapiro-Wilk test, while the overall plot suggests the approximate normal distribution of the residuals. The empirical density function curve of residuals (Figure 4) resembles the normal probability curve. Although the Shapiro-Wilk test suggests using a nonparametric test, it could have been a missed opportunity to use a more powerful parametric test. Fortunately in this case both nonparametric and parametric tests yield the same conclusions.

Differences in Computations of Shapiro-Wilk Statistics
As I mentioned before, comparing Shapiro-Wilk test in SAS 6.12 with that in SAS 8.2 using the same IQOL data, we see a slight difference in the W statistics (0.982009 in SAS Version 6.12, 0.987914 in SAS Version 8.2). In spite of a very small differences in W statistics, p-value for the tests are drastically different (p=0.219 in SAS Version 6.12, and p=0.0009 in SAS Version 8.2). So, there may be some changes in calculating p-values as well since the skewness of W alone cannot explain the difference.

DISCUSSION
In order to make the appropriate judgment in the distributional assumptions, we also need to look at the diagnostic plots that often provide the picture of the overall distribution along with the statistical tests. Combining graphical methods and test statistics will definitely improve our judgment on the normality of data. We can fairly easily automate the whole analysis process base on the result of single normal tests. However, it is a challenge from a programming point of view to incorporate such complicated visual assessment in our programs.

It has also been shown that small deviation in the test statistics had great effect on the single normality test. The SAS institute was consulted upon the observation of substantial differences in the p-values. According to the SAS institute, the reason for the difference is the improvement of algorithm to compute the W statistic.
FIGURE 1. Normal Q-Q plot of residuals from ANOVA model where percent changes in IEF is the dependent variable.

FIGURE 2. Empirical Density Function of residuals from ANOVA model where percent changes in IEF is the dependent variable (density estimates are obtained from PROC KDE).
FIGURE 3. Normal Q-Q plot of residuals from ANCOVA model where change in I-QOL is the dependent variable.

FIGURE 4. Empirical density function of residuals from ANCOVA model where change in IQOL is the dependent variable (density estimates are obtained from PROC KDE).
REFERENCES


ACKNOWLEDGEMENTS
I owe special thanks to Cathy Zhou, who provided valuable programming support for this paper. I would also like to thank Ilker Yalcin for his review and comments.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:
Guangbin Peng
Lilly Corporate Center
Indianapolis, Indiana 46285
U.S.A
Phone: 317 433 8445
Fax: 317 276 4789
Email: peng_guangbin@lilly.com