INTRODUCTION

For sample sizes up to \( n = 50 \), an ordered sample \((X_1, ..., X_n)\) may be tested for normality by a calculation involving a vector \((V_1, ..., V_n)\) that is tabulated in (1). If \( V \) and \( X \) are column vectors the test statistic, also tabulated in (1), is

\[
W = \frac{(V_i^2 - \bar{X})^2}{(X_i - \bar{X})^2}
\]

The distribution of \( W \) ranges from 0 to 1 where as \( W \) approaches 1.0 the distribution of \((X_1, ..., X_n)\) comes closer and closer to being Gaussian.

When dealing with a very large sample, \( N \) measurements, how does one proceed to test for normality since \( V \) for samples greater than 50 is not available? A common procedure is the chi squared test for deviations from an expected distribution. A more attractive but approximate method, outlined below, is the taking of a sample of \( n = 50 \) from the cross section of the large sample of \( N \) values. These 50 values may
then be tested by \((V_1\ldots V_{50})\) from the table in (1).

**METHOD**

The procedure for obtaining \(n = 50\) values from \(N\) values is:

The \(i^{th}\) ordered value of the sample \((X_1\ldots X_{50})\)
equals the \((N P_i + \frac{50-i}{49})^{th}\) ordered value of sample \((X_1\ldots X_N)\)
rounding to the nearest integral value.

The vector \(P = (P_1\ldots P_{50})\) is obtained from (2) using 50 percent ranks
with the tails of the distribution adjusted outward to give a standard
deveation for \((X_1\ldots X_{50})\) close to that of the large sample \((X_1\ldots X_N)\)
and with \(W\) close to 1. The vector \(P\) is:

\[
P = \{0.0048, 0.0322, 0.0531, 0.0729, 0.0928, 0.1126, 0.1325, 0.1524, 0.1722, 0.1921, 0.2119, \\
0.2318, 0.2517, 0.2715, 0.2914, 0.3113, 0.3311, 0.3510, 0.3709, 0.3907, 0.4106, 0.4305, \\
0.4503, 0.4702, 0.4901, 0.5099, 0.5298, 0.5497, 0.5695, 0.5894, 0.6093, 0.6291, 0.6490, \\
0.6689, 0.6887, 0.7086, 0.7285, 0.7483, 0.7682, 0.7880, 0.8079, 0.8278, 0.8476, 0.8675, \\
0.8873, 0.9072, 0.9270, 0.9469, 0.9678, 0.9952\}
\]

If \(N = 10,000\) the required set \((X_1\ldots X_{50})\) equals the (49th, 323rd,
532nd ... 9952nd) ordered values of the sample \((X_1\ldots X_{10,000})\).

The vector \((V_1\ldots V_{50}) = (-0.3751, \ldots, 0.0035, 0.0035, \ldots, 0.3751)\) and \(W\)
would be calculated as usual. From tables of the unit normal
distribution a cross-section sample of 50 yielded a sigma of 1.004
(for 50 degrees of freedom) and a \(W\) of 0.9996, both sufficiently close
to unity for general application.

SUMMARY

An alternate method has been outlined to test if a large sample is Gaussian in distribution. Instead of a chi squared test of fit a new statistic $W$ is evaluated using a cross-section sample of 50 from a much larger sample of data. If the large sample is at least 100, the technique yields reliable results which may be assessed for significance against tabulated percentiles of $W$.

JOHN SKORY
Math Statistician

REFERENCES
