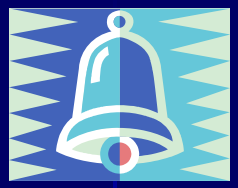
Using SAS to Assess Normality and Find the Optimal Transform

Brandy R. Sinco, Research Associate
University of Michigan
School of Social Work

Outline

- Skewness and Kurtosis
- Graphics in Proc Univariate
- Interpretation of QQ Plots
- Normality Statistics in Proc Univariate
- Box-Cox Method of Finding Optimal Normality Transform



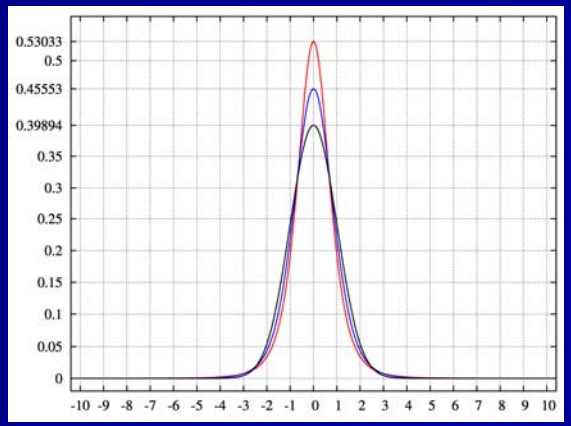
Normal Distribution, Skewness, Kurtosis

- X ~ Normal means that the density is:
- 2 parameters: μ = mean; σ^2 = variance

$$f(X) = \frac{\exp\left(\frac{-(X-\mu)}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}$$

- **μ_3, Skewness** = measure of asymmetry = $E(X - \mu)^3$
- Skewness is 0 for a normal distribution.

- **μ_4,** Kurtosis = meansure of peakedness = $E(X - \mu)^4$
- Kurtoisis is $3\sigma^4$ for a normal distribution and 3 for a standard normal distibution, N(0, 1) with μ = 0 and $\sigma^2 = 1$.



How SAS Computes Skewness and Kurtosis

Skewness	$$\mu_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s} \right)^3$$
Kurtosis **k_{n1}, k_{n2} are functions of n and close to 1 for large n.**	$$\mu_4 = k_{n1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s} \right)^4 - 3k_{n2}$$

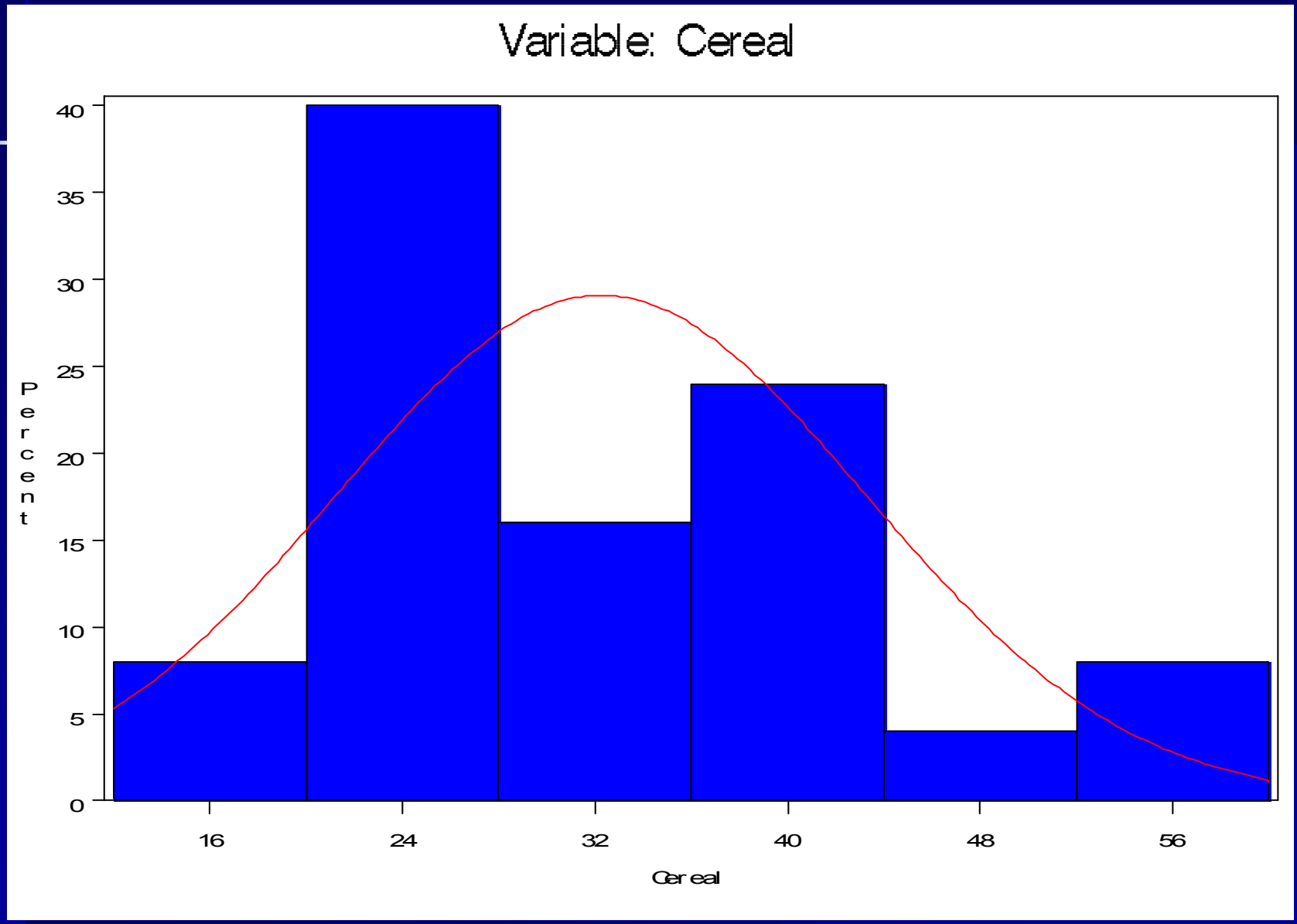
- SAS computes the standardized values of skewness and kurtosis for easy comparison to a standard normal distribution, N(0, 1)

Histogram and QQ Plot in Proc Univariate

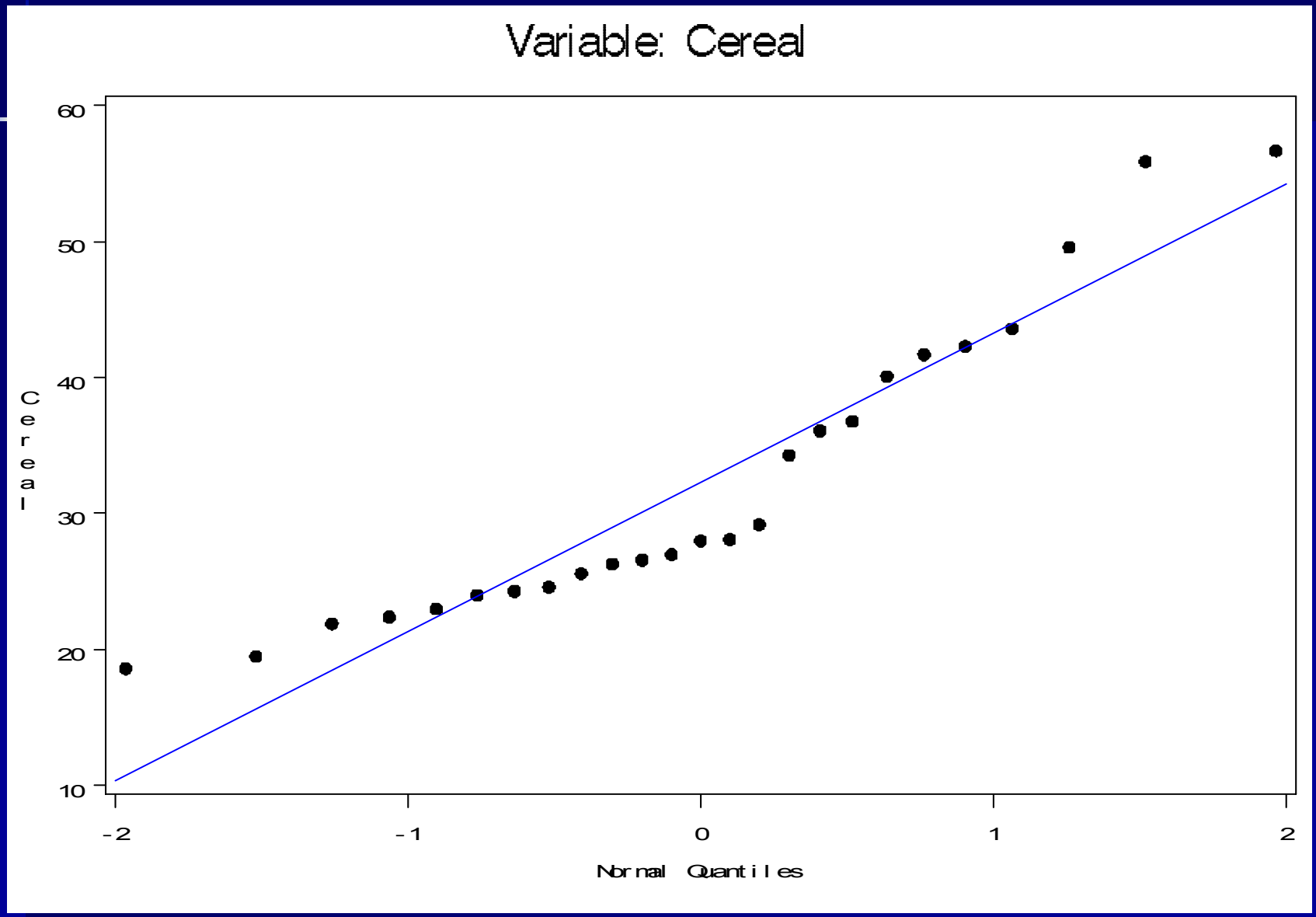
- Proc Univariate Data=SASUser.Protein;
- Var Cereal;
- Symbol1 V=Dot; /* SAS will use + without this symbol statement */
- Histogram / Normal(Mu=Est Sigma=Est Color=Red) CFILL=Blue;
- QQPlot / Normal(Mu=Est Sigma=Est L=1); Run;

- **Macro Conversion:**
- %Macro UniGraph(Vary, DatSet);
- Title "Variable: &VarY";
- Proc Univariate Data=&DatSet;
- Var &VarY;
- Symbol1 V=Dot;
- Histogram / Normal(Mu=Est Sigma=Est Color=Red) CFILL=Blue;
- QQPlot / Normal(Mu=Est Sigma=Est L=1); Run;
- %Mend UniGraph;

Sample Histogram From Proc Univariate



Sample QQ Plot From Proc Univariate
(skew = .93, kurt = -.07)



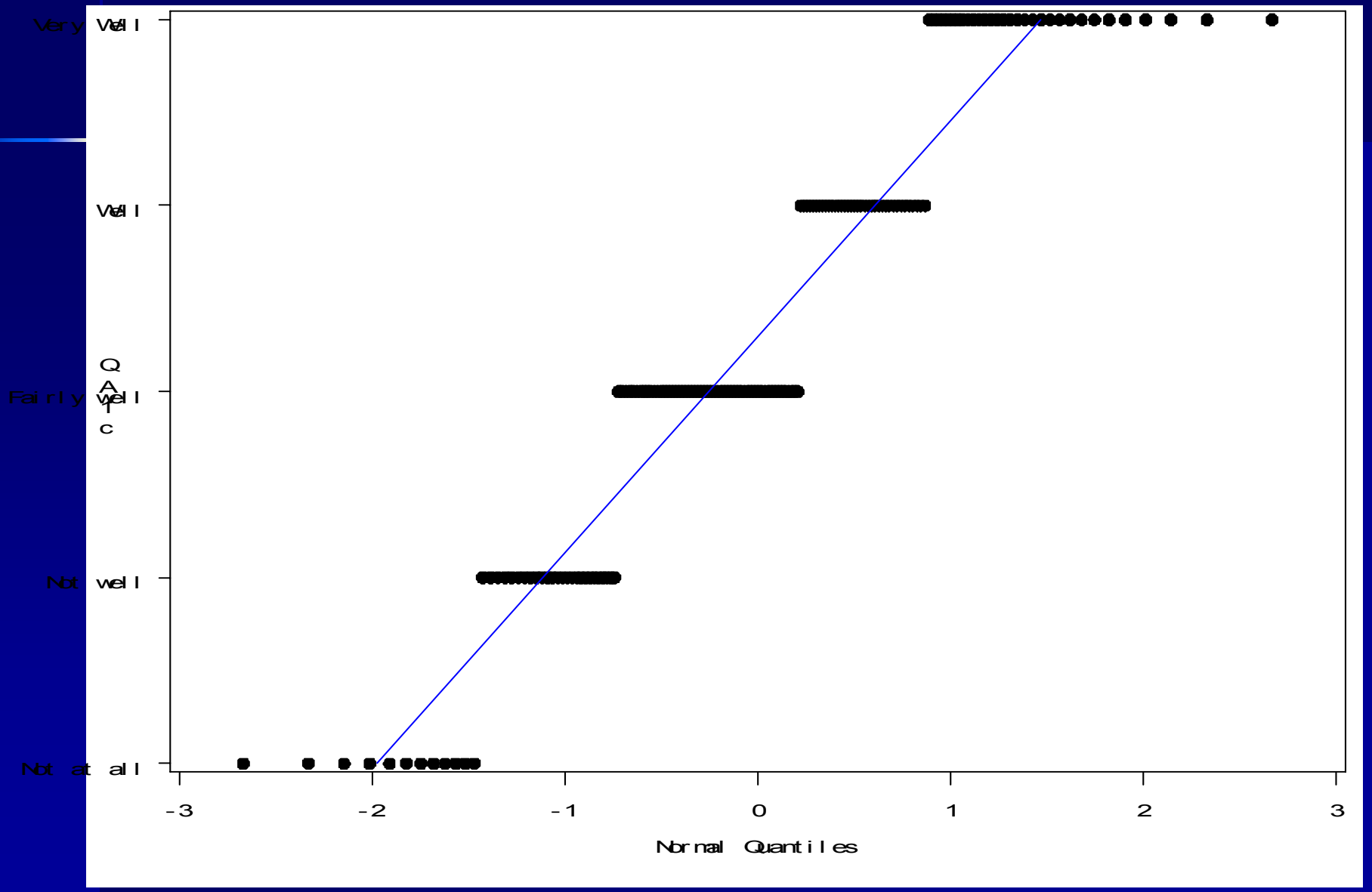
Theory Behind QQ Plot

- Plot ordered residuals, $e_{(k)}$ vs. rank-its r_k, which are theoretical values for a normal distribution
- Let Z = ordinate of standard normal distribution.
- Let s = standard deviation, α = percentile of standard normal.
- n = sample size.
- $r_k = sZ((k - .375)/(n + .25))$.
- **Note: SAS standardizes the normal quantiles.**

- Plot of $e_{(k)}$ vs. r_k should be a straight line.
- Sign of outliers: almost all points on the line, few far from line.
- Discrete or truncated data: staircase pattern

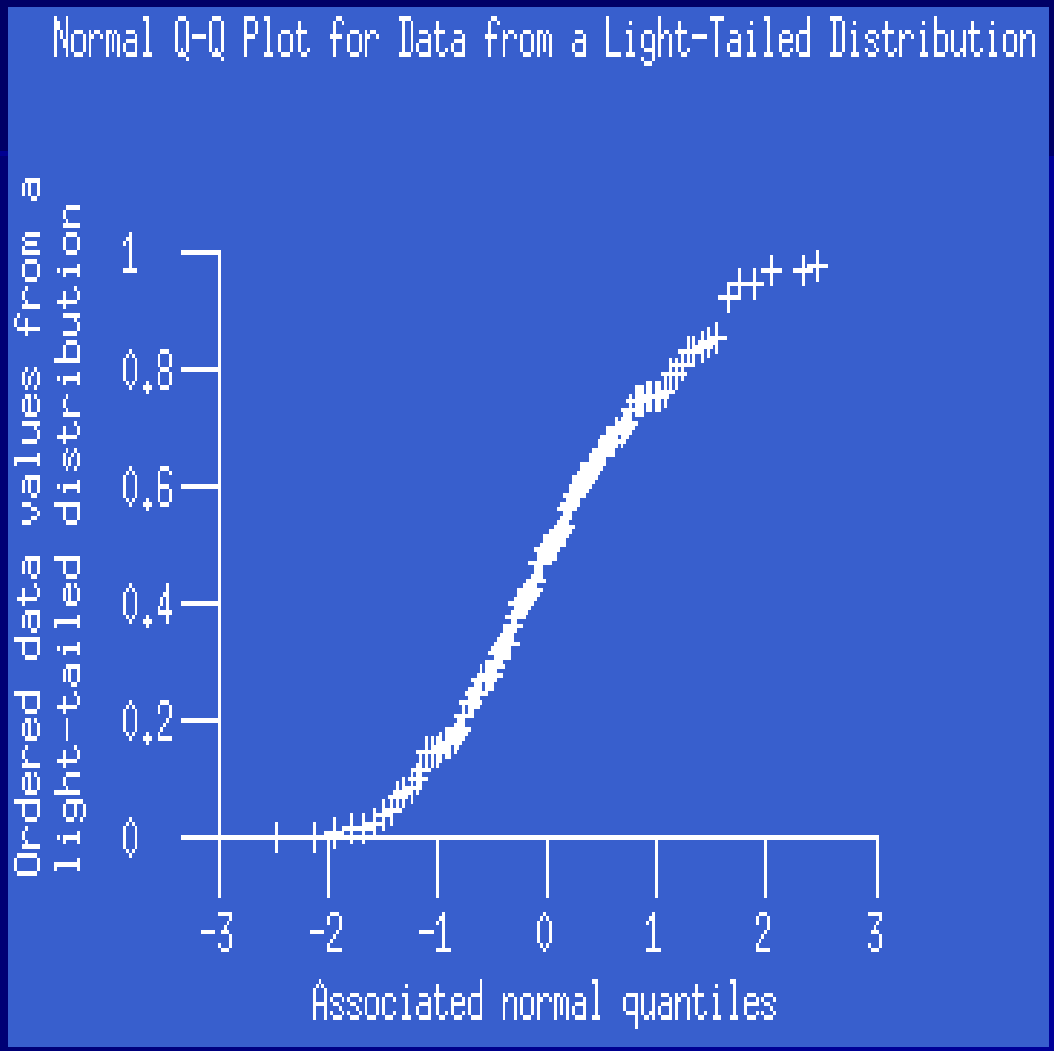
- References: Blom (1958), Chambers et al. (1983).

QQ Plot from Discrete or Truncated Variable



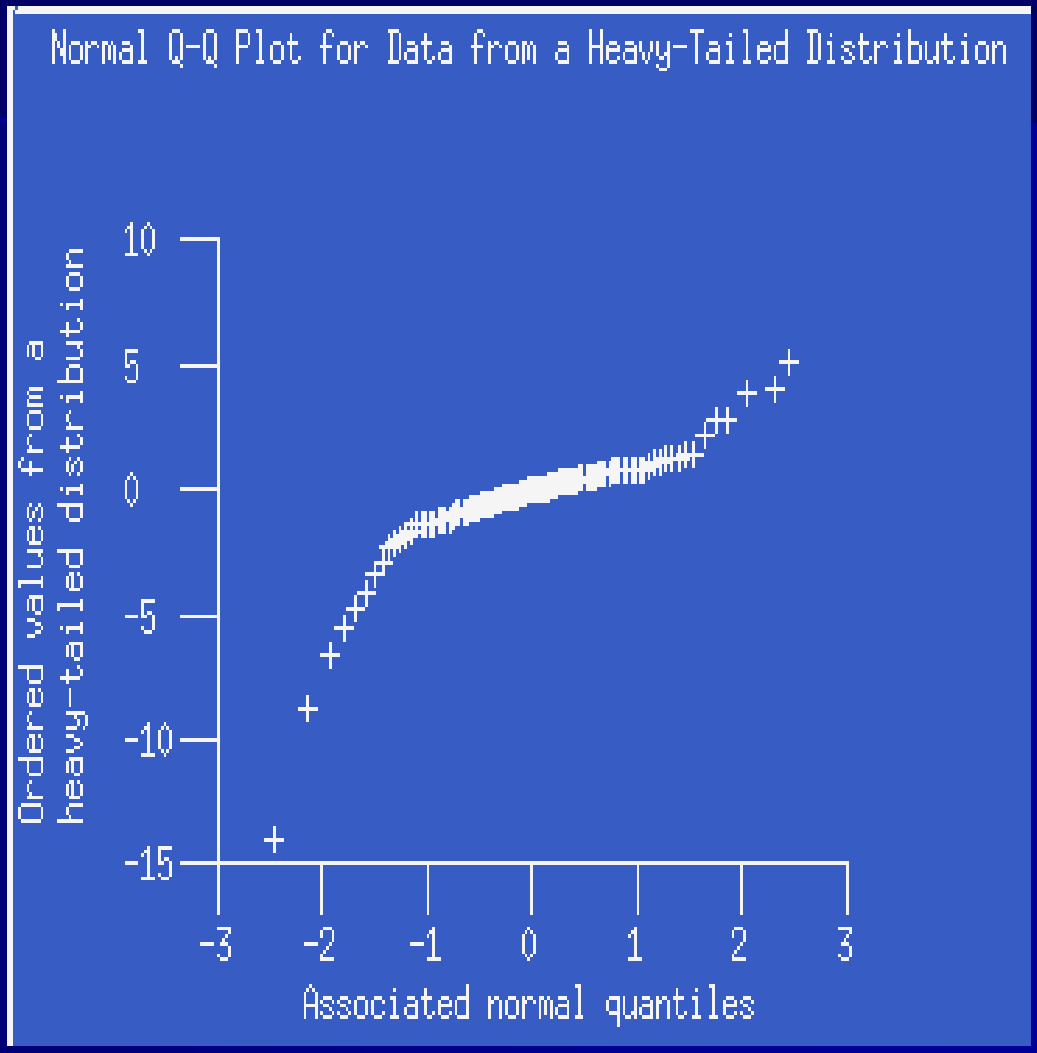
Light Tails (S Curve)

Right (upper) end of the normality plot bends below a hypothetical straight line passing through the main body of the X-Y values of the probability plot, while the left (lower) end bends above that line



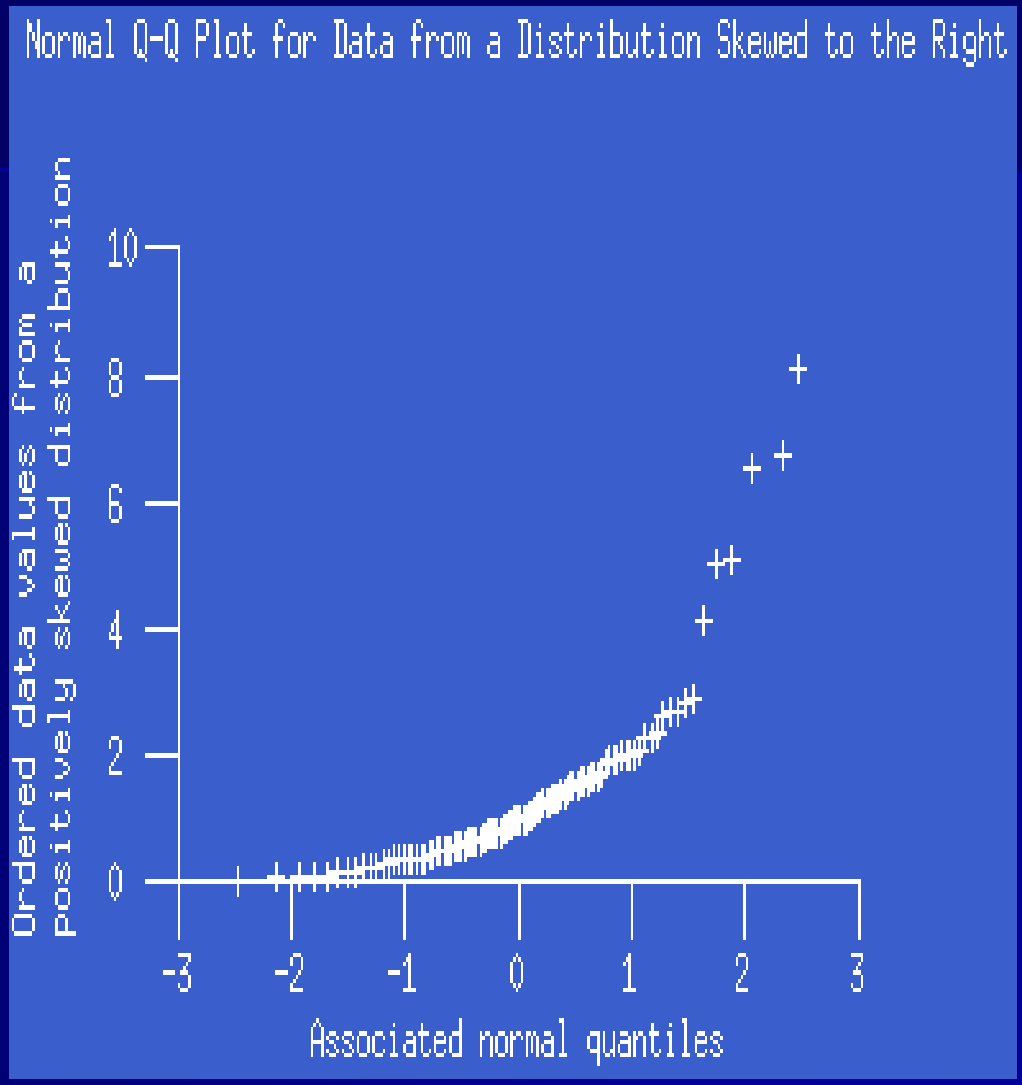
Heavy Tails (Reverse S Curve)

Right (upper) end of the normality plot bends above a hypothetical straight line passing through the main body of the X-Y values of the probability plot, while the left (lower) end bends below it



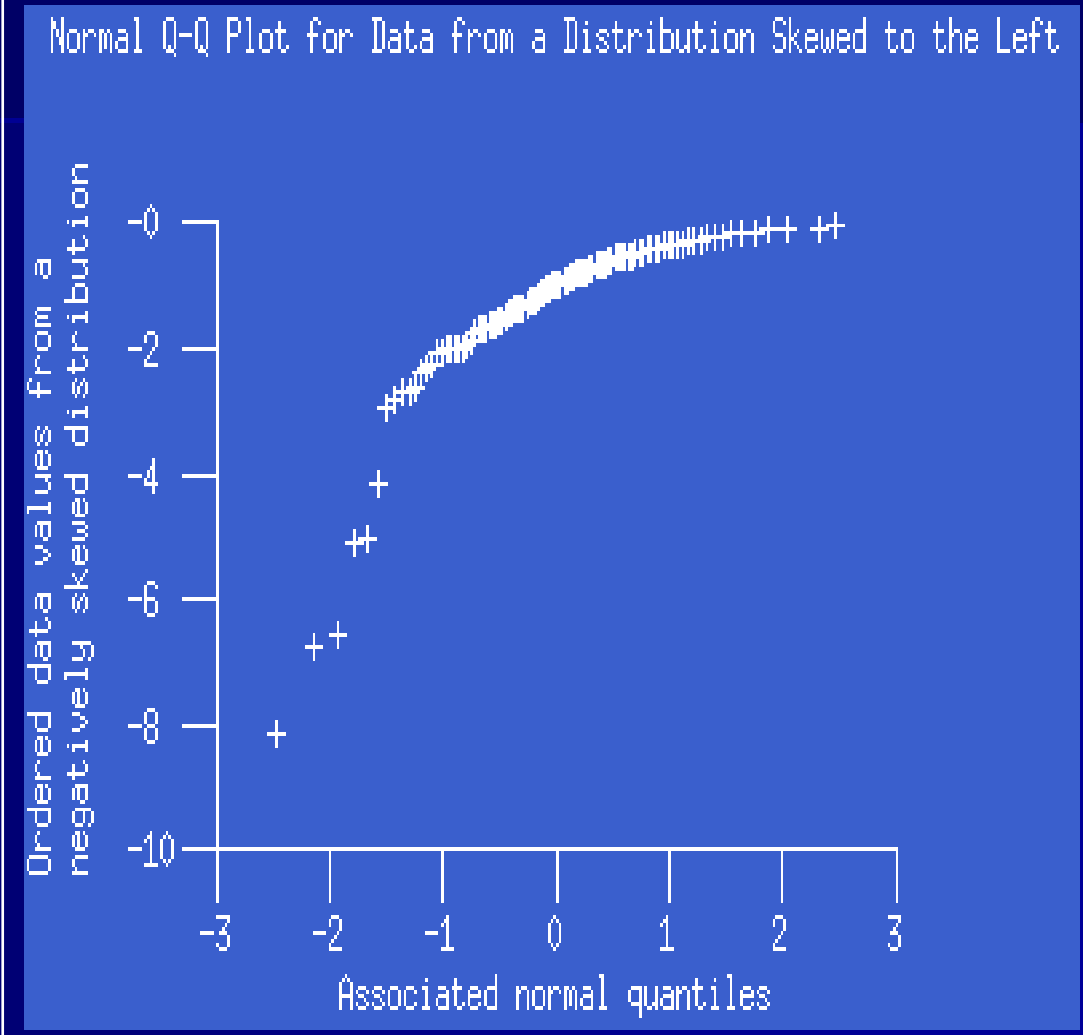
Skewed Right (Concave Upwards, Holds Water)

If both ends of the normality plot bend above a hypothetical straight line passing through the main body of the X-Y values of the probability plot, then the population distribution from which the data were sampled may be skewed to the right.



Skewed Left (Concave Downwards, Spills Water)

If both ends of the normality plot bend below a hypothetical straight line passing through the main body of the X-Y values of the probability plot, then the population distribution from which the data were sampled may be skewed to the left.



Normality Statistics from Proc Univariate

- SAS provides Shapiro-Wilk, Kolmogorv-Smirnov, Cramer-von Mises, and Anderson-Darling normality diagnostic statistics if Normal option is used.
- Proc Univariate Data=Test Normal;
- Var X;
- Run;

- Based on hypotheses:
- H_0: Distribution Is Normal. H_A: Distribution Is Not Normal.
- If data is exactly normal, p value will be large. I.E., if using $\alpha = .05$, $p > .05$.
- Often data is symmetrical enough for classical analysis techniques, such as regression, even if $p < .05$. Tests are overly conservative.
- Use these diagnostic statistic hand-in-hand with histogram and qq plot. If $p >> .05$, histogram looks normal, and qq plot follows a line, you can be confident that the data is very close to a normal distribution.
- However, if p is small and histogram and qq plot look close to normal, transforming the variable may not be necessary.

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)} \right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Shapiro-Wilk Test

- Published in 1965 by Samuel Shapiro and Martin Wilk.
- W statistic is ratio of order statistics over sample variance.
- $X_{(i)}$ = ith order statistic.
- a_i = parameter calculated from expected values of order statistics of a standard normal, N(0, 1).
- Denominator = estimated variance.

- $0 \leq W \leq 1$. Values close to 1 indicate normality; small values indicate non-normality.
- Valid for $n \leq 2000$.

Empirical Distribution Function (EDF), $F_n(x)$

- Shapiro-Wilk test is in a class by itself.
- Kolmogorv-Smirnov, Cramer-von Mises, and Anderson-Darling statistics are based on the empirical distribution function.

- EDF has a staircase shape
- $X_{(i)}$ = ith order statistic.
- $F(x)$ = Cumulative Distribution Function = $Pr(X \leq x)$
- $F_n(x)$ = Empirical Distribution Function for Sample of Size n.

- $F_n(x) = 0$ if $x < X_{(1)}$
- $F_n(x) = i/n$ if $X_{(i)} \leq x < X_{(i+1)}$
- $F_n(x) = 1$ if $x \geq X_{(n)}$

Kolmogorov-Smirnov Test

- Introduced by Kolmogorov in 1933; asymptotic tabulations by Smirnov in 1948.
- Let x = variable being evaluated.
- Let F(x) = cumulative distribution function for for a normal random variable with mean = average(x) and variance = var(x).
- Let $F_n(x)$ = empirical distribution function of x.

- $D = \sup_x | F_n(x) – F(x)|$
- Kolmogorov-Smirnov test statistic based on maximum distance between empirical distribution function, $F_n(x)$, and cumulative distribution function for a normal random variable with the mean and variance of x.
- Large sample size, n > 2000, needed.
- More sensitive to departures from normality near the center than in the tails.

Anderson-Darling and Cramér-von Mises Tests

- Both test statistics are based on $(F(x) - F_n(x))^2$.
- Anderson-Darling uses a weighting algorithm.
- Some analysts think these statistics, particularly Anderson-Darling, estimate the distribution better than Kolmogorov-Smirnov.
- References:

- Conover, W. J. (1999), Practical Nonparametric Statistics, Third Edition, New York: John Wiley & Sons, Inc.
- D'Agostino, R. B. and Stephens, M. (1986), Goodness-of-Fit Techniques, New York: Marcel Dekker, Inc.

Box-Cox Transform

- Find λ, such that Y^{λ} is closest to normality.
- Box, G.E.P. and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistics Society*, B-26, 211 - 252.
- Uses method of maximum likelihood estimation.
- Can be used for a regression model, such $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, or individual Y variable.
- $\lambda = 0$ corresponds to log transform.
- $\lambda = 1$ corresponds to original Y variable.
- Best to round λ to closest multiple of .5 or .25. I.E., if $\lambda = .348$, choose .5 or .25.
- Transforming data back and forth for estimates is awkward if λ is an obscure number and usually doesn't help the model.

Box-Cox With SAS Proc TransReg

- To use original Y value:
- Data TestData1;
- Set TestData;
- One=1;
- Run;

- Proc TransReg Data=TestData1 NOZ;
- /* NOZ = Intercept Only Model, $Y = \beta_0 + \varepsilon$ */
- Model BoxCox(Y) = Identity(One);
- Run;

- To use TransReg on regression model, to normalize residuals:
- Proc TransReg Data=Test; /* Don't use NOZ option */
- Model BoxCox(Y) = Identity(X1 X2);

SAS Output from Proc TransReg

- (<) Maximum Likelihood Estimate (MLE) of λ
- (+) Convenient value of λ selects 1 or 0 if inside confidence interval, closest integer, or closest multiple of 0.5.
- (*) 95% confidence interval for λ.

Lambda	R-Square	Log Like
-2.00	0.01	-64.5625 *
-1.50	0.01	-64.2914 *
-1.00	0.01	-64.2035 <
-0.50	0.01	-64.3046 *
0.00	0.02	-64.5999 *
0.50	0.02	-65.0951 *
1.00 +	0.02	-65.7949 *
1.50	0.02	-66.7036

- Default λ range is -3 to 3 by .25

Transformations and Skewness

- Tukey's Transformation Ladder (1977).

- To eliminate right skewness, transform to lower powers.
- $X^{.5}$, $X^{.25}$, ln(X).
- Recall that $\lambda = 0$ corresponds to ln(X)

- To eliminate left skewness, tranform to higher powers.
- X^2, X^3, exp(X), etc.

Summary: Assessing Normality with SAS

- Skewness and Kurtosis from Proc Univariate.
- SAS reports standardized values for distibution with mean 0 and variance 1.
- No need to standardize data; SAS does it automatically.

- Histograms and QQ Plots.

- Normal diagnostic statistics.
- Non-significant p values indicate normality, because null hypothesis is that data is normal.
- Use them to confirm normality of plots.
- Don’t be alarmed by small p values.

- Box-Cox Transformation with Proc TransReg.

Contact Information

- Brandy R. Sinco
- University of Michigan School of Social Work
- 1080 S. University St.
- Box 183
- Ann Arbor, MI 48109-1106

- Phone: 734-763-7784

- E-Mail: brsinco@umich.edu