

SAS Tools for Assessing Multivariate Normality

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

Acknowledgement of Dr. Phil Chapman

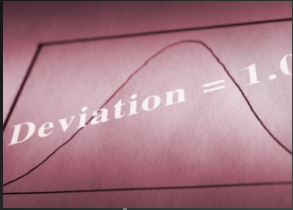
- The material in my presentation has been taken from my masters project on Structural Equation Modeling.
- Although the content was created by me, Dr. Chapman deserves acknowledgment for reviewing my writing for accuracy.
- Dr. Chapman is professor of Statistics at Colorado State University, Fort Collins, CO.

Outline

- What is multivariate normality?
- Statistical distance.
- Histogram and QQ Plot
- Ellipse plots for two variables.
- Mardia's tests for multivariate skewness and kurtosis.

Why is multivariate normality important?

- “We begin this multivariate statistics class by discussing multivariate normality because the rest of this class depends on it”. Dr. Donald Estep, Professor of Mathematics and Statistics at Colorado State University.
- Principle Components, Factor Analysis, Discriminant Analysis, Cluster Analysis.
- Important assumption for Structural Equation Modeling.



Univariate Normal

- $X \sim$ Univariate Normal: $\mu =$ mean; $\sigma^2 =$ variance

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

- $Z = (X - \mu)/\sigma \sim N(0, 1)$, standardized normal variable.
- $Z^2 \sim$ Chi-Square(1).
- Skewness = measure of symmetry = 0.
- Kurtosis = measure of peakedness = $E(X - \mu)^4$
- Kurtosis is $3\sigma^4$ for a normal distribution ; 3 for a standard normal distribution, $N(0, 1)$ with $\mu = 0$ and $\sigma^2 = 1$.



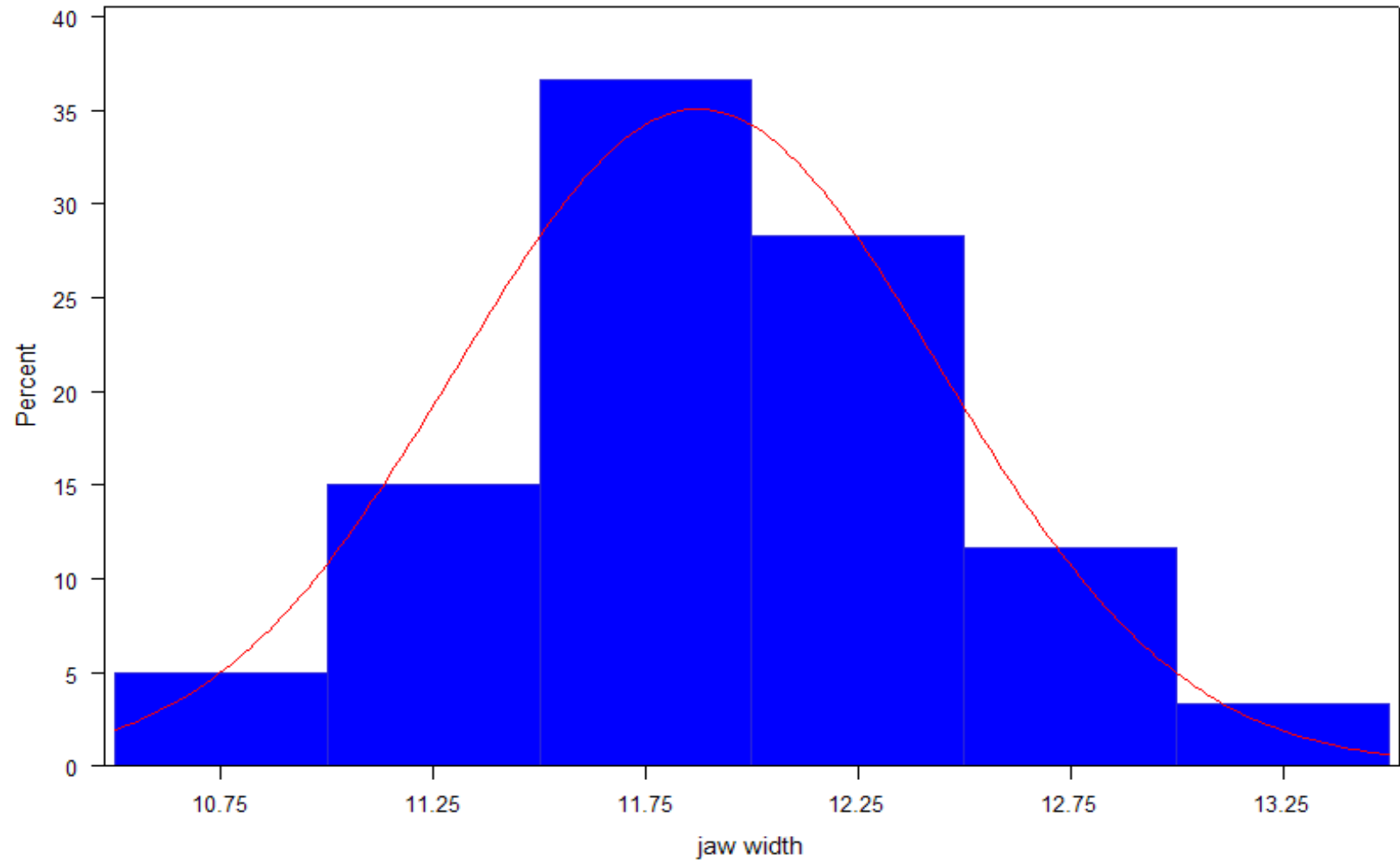
Multivariate Normal

- $W \sim$ Multivariate normal , $W = [W_1 \ W_2 \ \dots \ W_p]^t$
- with mean $[\mu_1 \ \mu_2 \ \dots \ \mu_p]^t$ and variance matrix Σ

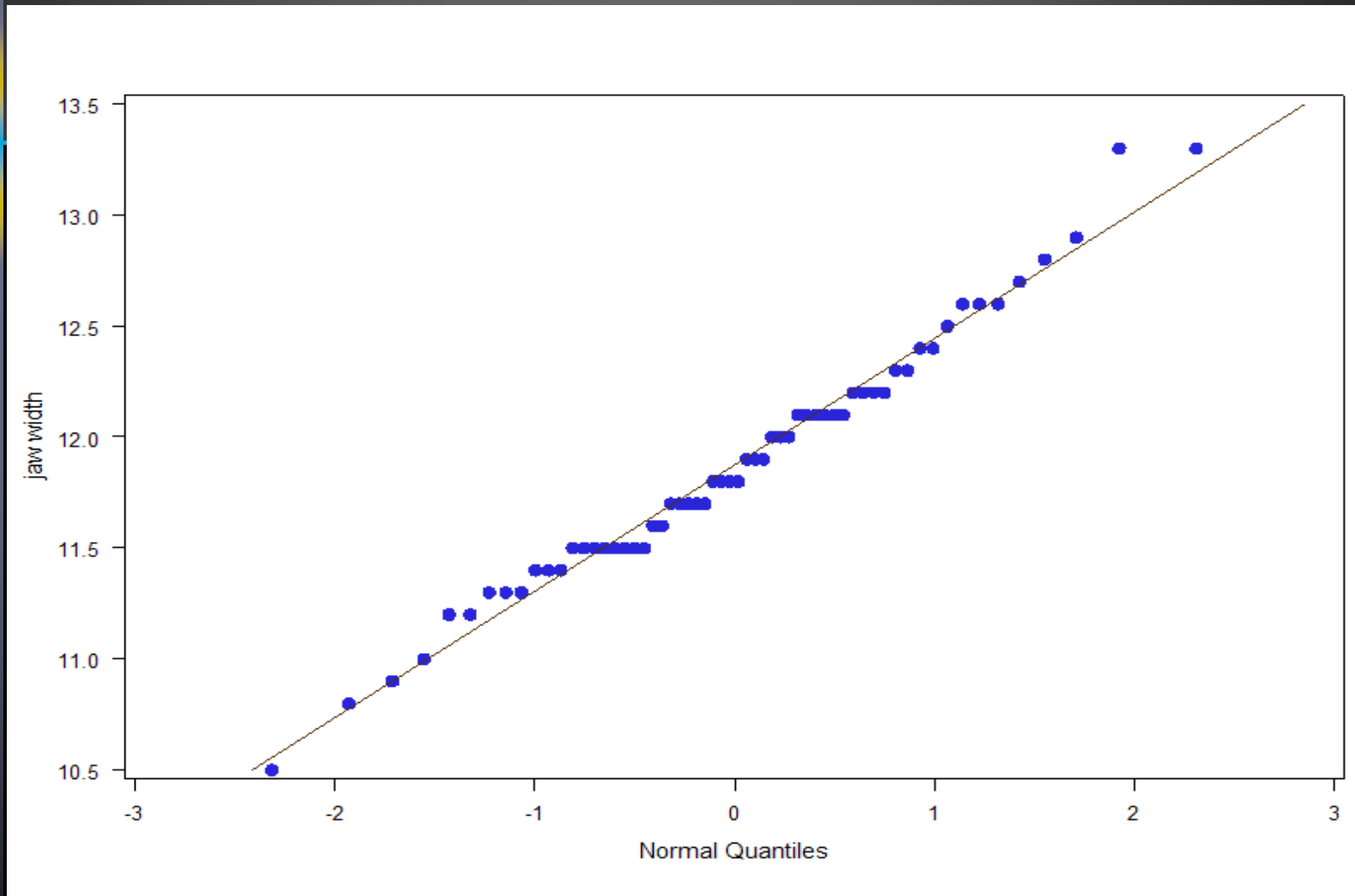
$$f(W) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(W - \mu)^T \Sigma^{-1} (W - \mu)\right)$$

- Σ must be positive definite (positive determinant).
- $Y = (W - \mu)^t \Sigma^{-1} (W - \mu) \sim$ Chi-Square(p).
- QQ Plot and Histogram for a single variable, X , compared to $N(0,1)$ or Chi-Square(1).
- For a set of variables, W , compare to Chi-Square(p)

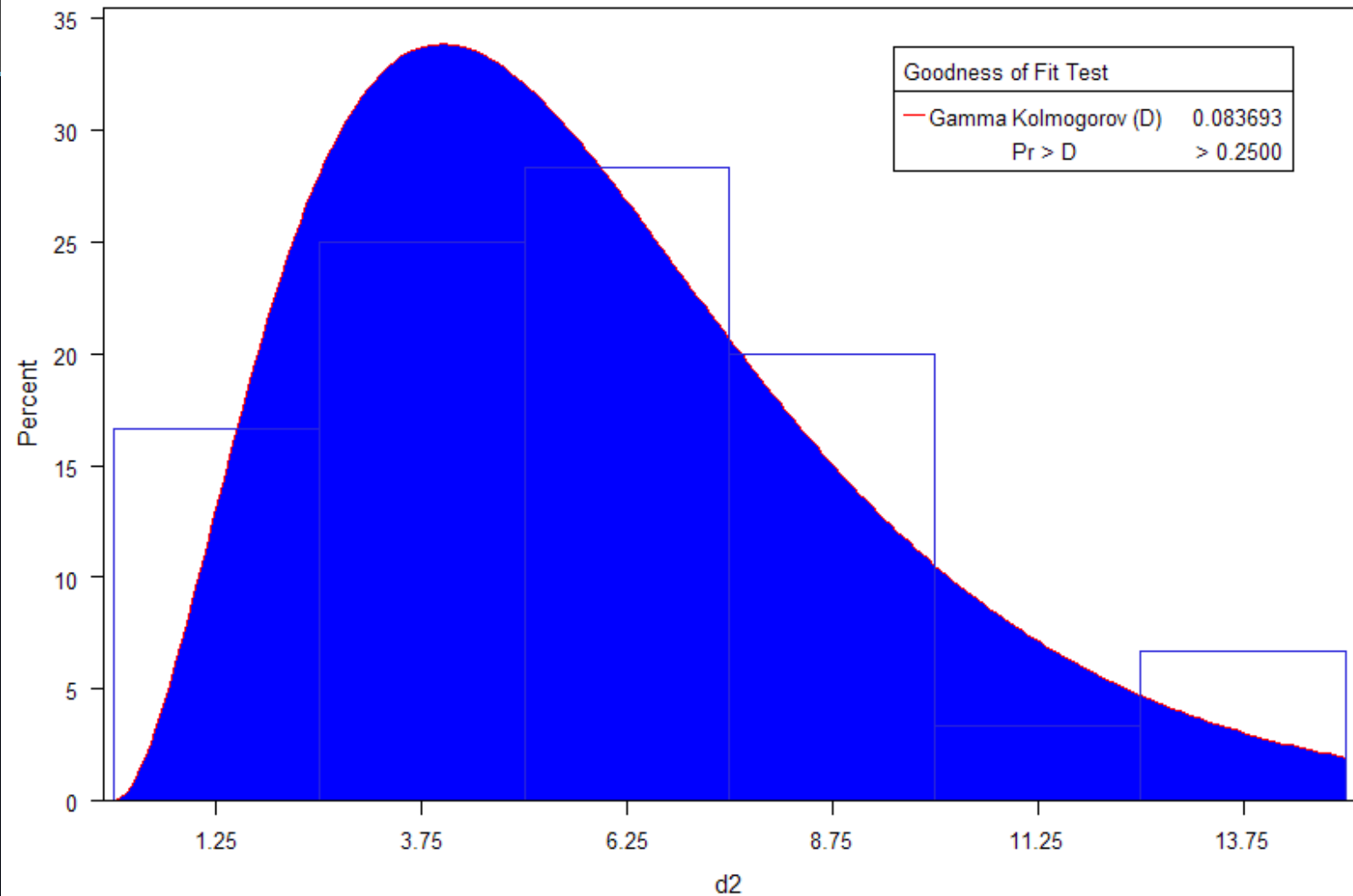
Sample Histogram From Proc Univariate



Sample QQ Plot From Proc Univariate (skewness= 0.28, kurtosis = 0.30)



Multivariate Histogram Compared to Chi-Square(p)



Univariate Histogram and QQ Plot in Proc Univariate

- Proc Univariate Data = Football
- Var Jaw;
- 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;

- Skewness and Kurtosis from Proc Univariate will be for the standardized variable (subtract mean and divide by the standard deviation).
- Want skewness to be close to 0.
- Proc Univariate subtracts 3 from the kurtosis.
- Want kurtosis to be close to 0.

Multivariate QQ Plot and Histogram – Step 1: Distance

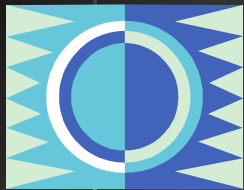
- First, compute the Mahalanobis distance for the standardized variables.
- `/* Use the std option to direct SAS to use the standardized correlation matrix */`
- `proc princomp data=Football std out=outfootball;`
- `var WDIM CIRCUM FBEYE EYEHD EARHD JAW;`
- `run;`
- `data mahalanobis;`
- `set outfootball;`
- `d2=uss(of prin:);`
- `run;`

Multivariate QQ Plot & Histogram – Step 2: Chi-Square

- Chi-squared distribution with 6 degrees of freedom (6 X's).
- Chi-squared distribution is a special case of gamma distribution.
- ALPHA = df/2, SIGMA = 2, THETA=0; (df = a * σ)
- Proc Univariate Data=Mahalanobis;
- Var d2;
- Histogram / Gamma(Alpha=3 Sigma=2 Theta=0 Fill Color=Red) CFill=Blue Name="Football";
- Inset Gamma(KSD KSDPval) / Header='Goodness of Fit Test' Position=(95,95) RefPoint=TR;
- QQPlot/ Gamma(Alpha=3 Sigma=2 Theta=0 L=1) PctlMinor PCTLSCALE Name="QQ Plot";
- Run;

Kolmogorov-Smirnov Goodness of Fit Statistic

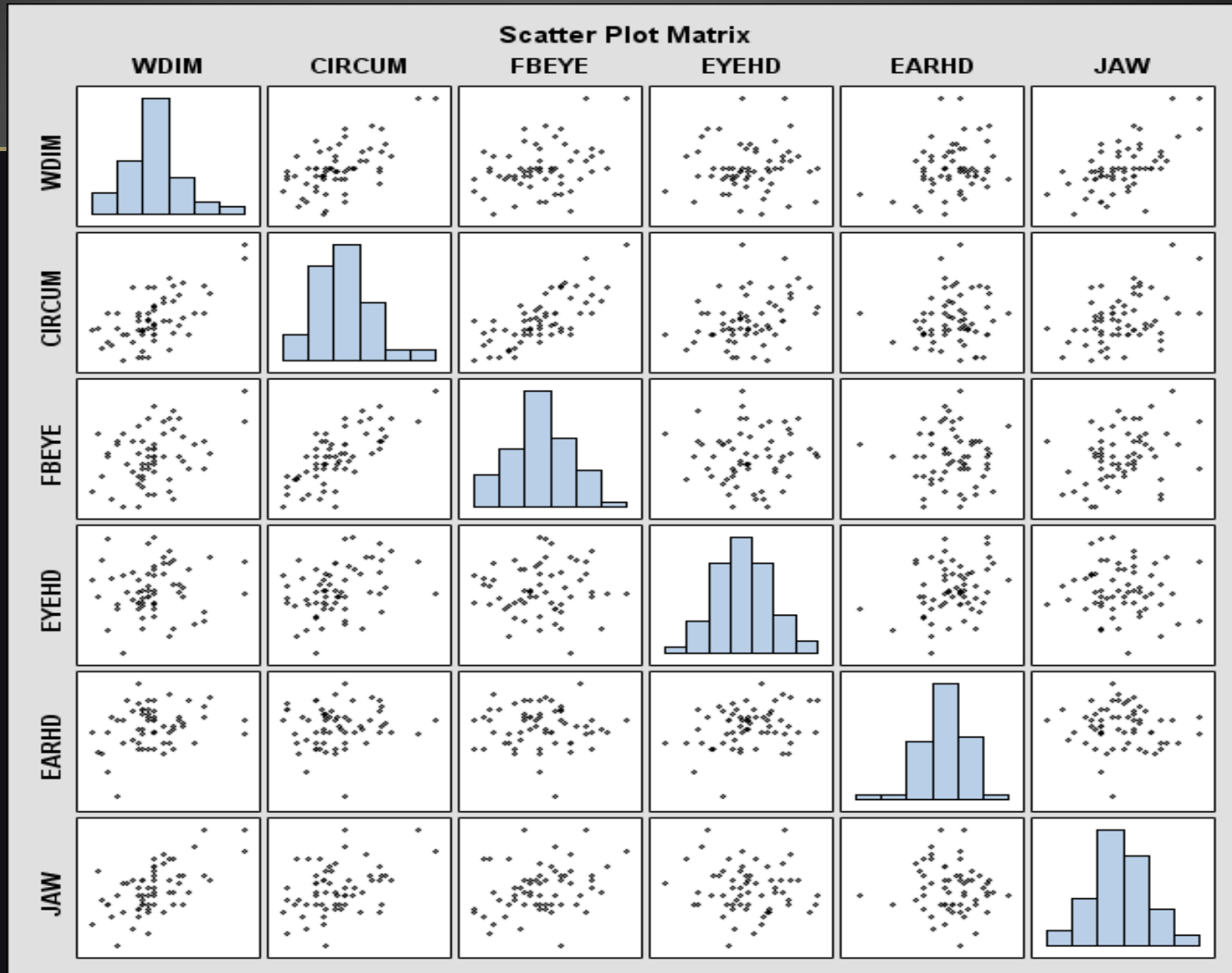
- Based on a comparison between the empirical and hypothesized cumulative distribution functions (Kolmogorov, 1933; Smirnov, 1948).
-
- H_0 : Distribution is $N(0, 1)$ for univariate or $\text{Chi-Sq}(p)$ for multivariate
- H_A : Distribution differs from hypothesized distribution.
- If data is exactly normal, p value will be large. I.E., if using $\alpha = .05$, $p > .05$.
- While a non-significant p value indicates normality, this test is overly conservative and rejects too easily.
- Kolmogorov-Smirnov unreliable for $n < 2000$ (Peng and Lilly, 2004, SAS Institute, 2011b). For univariate normal test, better to use the Shapiro-Wilk statistics for $n \leq 2000$.
- Use this diagnostic statistic hand-in-hand with histogram and qq plot. If $p \gg .05$, histogram looks normal, and qq plot follows a line, you can be confident that the data is very close to a normal distribution.



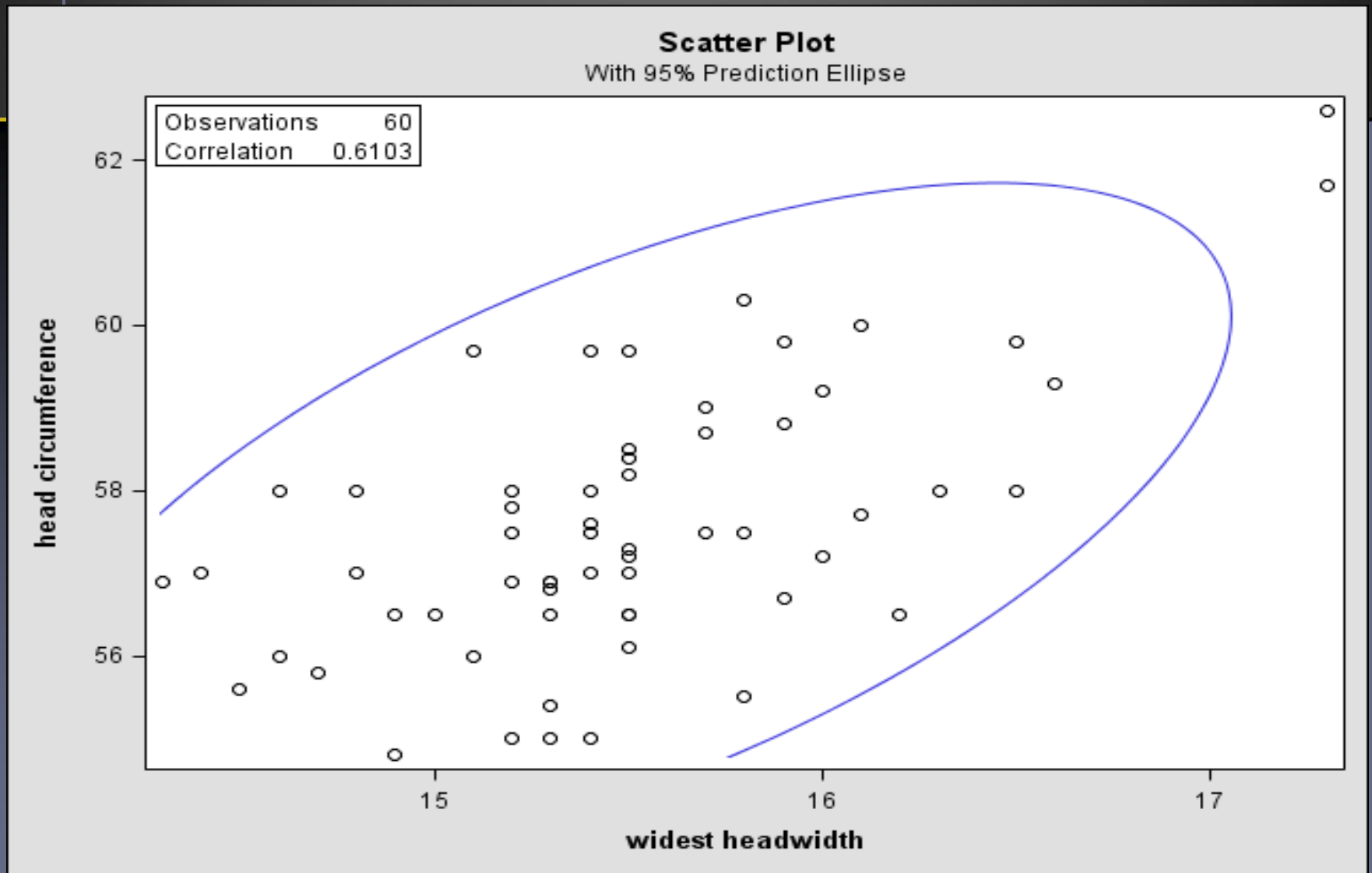
Bivariate Normality

- For multivariate analysis techniques, concerned about estimating mean and covariance matrix. Interested in univariate, bivariate, and multivariate (all variables together) normality.
- For bivariate normality, scatter plot of W1 and W2 should be within an ellipse.
- For one variable, confidence interval is a line segment.
- For two normal variables, confidence interval is an ellipse.
- ods graphics on;
- Proc Corr cov Data=football plots=(matrix(nvar=6 hist) scatter(alpha=.05));
- var WDIM CIRCUM FBEYE EYEHD EARHD JAW;
- Run;
- ods graphics off;
- /* hist option puts histograms on diagonal of pairs plots */

Scatter Plot Matrix from Proc Corr



Ellipse Plot from Proc Corr





Mardia's Multivariate Skewness and Kurtosis

- K.V. Mardia (1970) showed that a p -dimensional standard normal variable will have kurtosis $p(p + 2)$. Note that if $p = 1$, kurtosis = 3.

- Multivariate Skewness.
$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left[(w_i - \bar{w})^T S^{-1} (w_j - \bar{w}) \right]^3$$

- $nb_{1,p}/6 \sim \text{Chi-Square}(p(p + 1)(p + 2)/6)$

- Multivariate Kurtosis.
$$b_{2,p} = \frac{1}{n^2} \sum_{i=1}^n \left[(w_i - \bar{w})^T S^{-1} (w_i - \bar{w}) \right]^2$$
- $b_{2,p} \sim N(p(p + 2), 8p(p + 2)/n)$



Kurtosis in Proc CALIS

Proc CALIS is used for Structural Equation Modeling (SEM) .

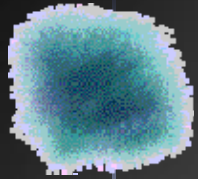
- Multivariate Kurtosis is evaluated based on Mardia's statistics.
- To get Kurtosis, use the "Kurtosis" option on the model statement.
- While Proc CALIS will give a variety of Kurtosis measures, the most important one is the "Normalized Multivariate Kurtosis".
- If the data is multivariate normal, the "Normalized Multivariate Kurtosis" will be between +/- 1.96, because the distribution of the test statistic will be $N(0, 1)$.



Mardia's Skewness and Kurtosis in Proc Model

Use the normal option on the "fit" statement to get Mardia's skewness and kurtosis, along with p values.

- `proc model data=semdata;`
- `Pre_H1c = E1;`
- `Age_BLIW = E2;`
- `TotalClasses = E3;`
- `FHADrVisitsCat = E4;`
- `Post_H1c = E5;`
- `fit Pre_H1c Age_BLIW TotalClasses FHADrVisitsCat Post_H1c /
normal ;`
- `run;`
- `quit;`



Summary

- Interested in univariate, bivariate, multivariate (all variables together) normality for classical multivariate procedures.
- Proc Univariate qq plot and histogram.
- For multivariate qq plot and histogram, use the gamma distribution with $\alpha = (\text{num variables})/2$ and $\sigma = 2$; $df = \text{num variables}$.
- For bivariate ellipse plots, use Proc Corr with scatter(alpha =) option.
- Proc CALIS will display Mardia's Kurtosis; most important measure is the "Normalized Multivariate Kurtosis", which should be ≤ 1.96 in absolute value.
- Proc Model will display multivariate skewness and kurtosis with p values.

References

- Johnson, R. A. and Wichern, D. W. (2007). Applied Multivariate Statistical Analysis, 6th ed. New Jersey: Prentice Hall.
- Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. Giorn. 1st. Ital. Attuari, 83-91.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519-530.
- Peng, G. and Lilly, E. (2004). Testing Normality of Data using SAS, Paper PO04. SUGI Conference Proceedings, 1-6.

- SAS Institute. (2011). The CALIS procedure.
- SAS Institute. (2011). The Model procedure.
- SAS Institute. (2011). The PrinComp procedure.
- SAS Institute. (2011). The Univariate procedure.

- Smirnov, N. V. (1948). Tables for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 19, 279-281.



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