

# Generalized Variance Multivariate Normal Distribution

Lecture 4

September 14, 2005

Multivariate Analysis

Overview

Last Time

Today's Lecture

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Matrices and vectors.
  - ◆ Eigenvalues.
  - ◆ Eigenvectors.
  - ◆ Determinants.
- Basic descriptive statistics using matrices:
  - ◆ Mean vectors.
  - ◆ Covariance Matrices.
  - ◆ Correlation Matrices.

# Today's Lecture

Overview

Last Time

Today's Lecture

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Generalized Variance (Chapter 3, Section 4).
- Multivariate Normal Distribution (Chapter 4).

# Sample Covariance Matrix

Overview

Generalized Variance

Generalized Sample Variance  
Generalized Sample Variance  
With  $\mathbf{R}$   
Total Sample Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Recall the sample covariance matrix:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} (\mathbf{X} - \mathbf{1}\bar{x}')' (\mathbf{X} - \mathbf{1}\bar{x}').$$

- The overall sample covariance matrix gives a picture of the covariation between each variable in the sample.
- A single-number summary of this matrix can be provided by the generalized variance.
- Although the generalized variance is not used frequently, you will see in later slides that this value is part of the multivariate normal distribution.

# Generalized Sample Variance

Overview

Generalized Variance

Generalized Sample Variance

Generalized Sample Variance

With  $\mathbf{R}$

Total Sample Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Generalized Sample Variance is computed by  $|\mathbf{S}|$  (the determinant of the sample covariance matrix).
- To begin, imagine a multidimensional cube (an ellipsoid) that represents the end points of all the column vectors in the sample matrix  $\mathbf{X}$ .
  - ◆ Covariance matrix describes the overall spread of that shape in each direction (if we could plot it).
  - ◆ The equation  $(\mathbf{x} - \bar{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = a^2$  describes an equation where all points are equal distant from the mean  $\bar{\mathbf{x}}$  (i.e., it will form an ellipsoid).
  - ◆ That ellipsoid will have axes proportional to the square roots of the eigenvalues of  $\mathbf{S}$ .
  - ◆ The volume of that ellipsoid is equal to  $|\mathbf{S}|^{1/2}$  (notice what happens if we have a zero eigenvalue?)

# Generalized Sample Variance With R

Overview

Generalized Variance

Generalized Sample Variance

Generalized Sample Variance

With R

Total Sample Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- The generalized sample variance is dependent upon the scale of the variables in the sample.
- Because of the scale of these variables, the generalized sample variance can be hard to interpret (much like variances and covariances).
  - ◆ The scale of a single variable may have a disproportionate impact on the generalized variance (e.g., your sample has the variables of GPA and income in US dollars).

# Generalized Sample Variance With R

Overview

Generalized Variance

Generalized Sample Variance

Generalized Sample Variance

With R

Total Sample Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Rather than using  $|\mathbf{S}|$ , the sample correlation matrix can be used  $|\mathbf{R}|$ .
- The interpretation of the GSV is the same - the volume of the ellipsoid that is formed by the standardized variables in the sample.
- The difference in magnitude is proportional to the product of the variances of the variables in the sample.
- SAS Example #1...

# Total Sample Variance

Overview

Generalized Variance

Generalized Sample Variance

Generalized Sample Variance

With  $\mathbf{R}$

Total Sample Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Another way to characterize the sample variance is with the total variance.
- Total variance equals  $tr(\mathbf{S}) = s_{11} + s_{22} + \dots + s_{pp}$ .
- Describes the variability of the data without taking in to account the covariances.
- Like generalized variance, the total sample variance reflects the overall spread of the data.
- Many multivariate techniques refer use total sample variance in computation of variance accounted for.

# Multivariate Normal Distribution

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- The generalization of the well-known normal distribution to multiple variables is called the multivariate normal distribution (MVN).
- Many multivariate techniques rely on this distribution in some manner.
- Although real data may never come from a true MVN, the MVN provides a robust approximation, and has many nice mathematical properties.
- Furthermore, because of the central limit theorem, many multivariate statistics converge to the MVN distribution as the sample size increases.

# Univariate Normal Distribution

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- The univariate normal distribution function is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2}$$

- The mean is  $\mu$ .
- The variance is  $\sigma^2$ .
- The standard deviation is  $\sigma$ .
- Standard notation for normal distributions is  $N(\mu, \sigma^2)$ , which will be extended for the MVN distribution.

# Univariate Normal Distribution

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

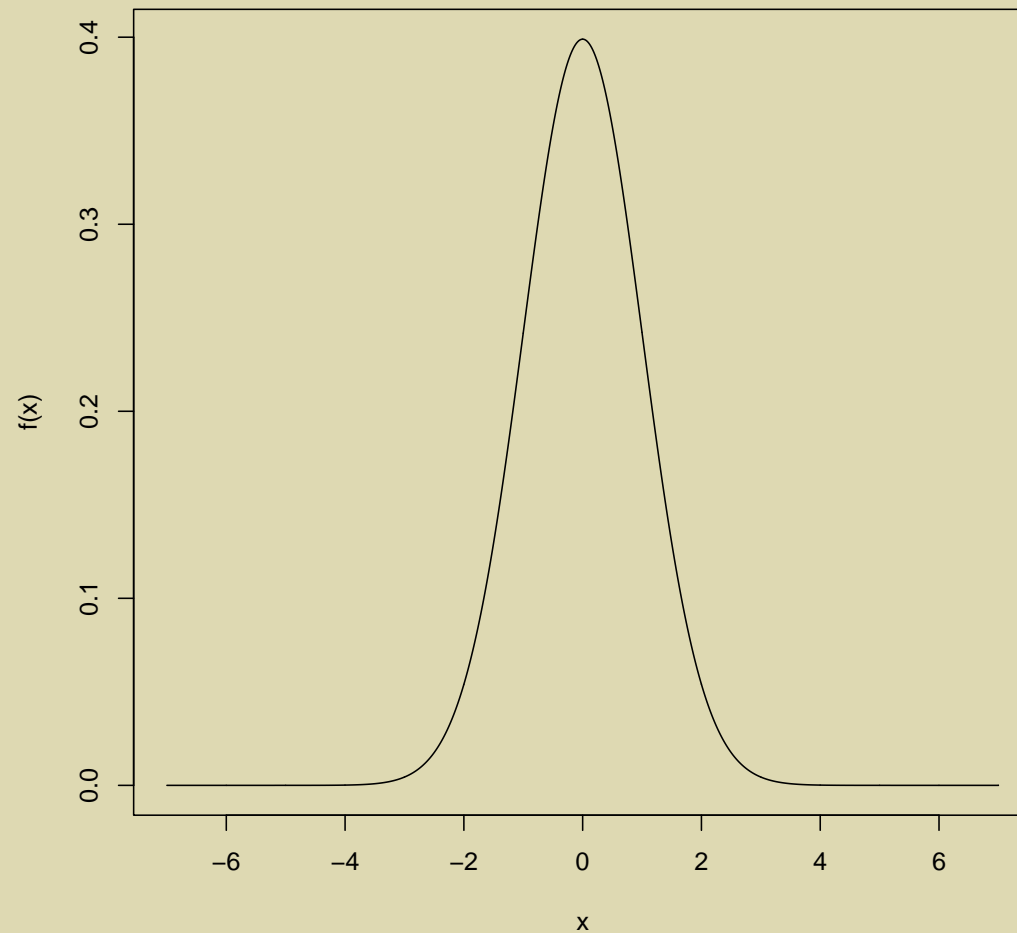
Outliers

Transformations

Wrapping Up

$$N(0, 1)$$

Univariate Normal Distribution



# Univariate Normal Distribution

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

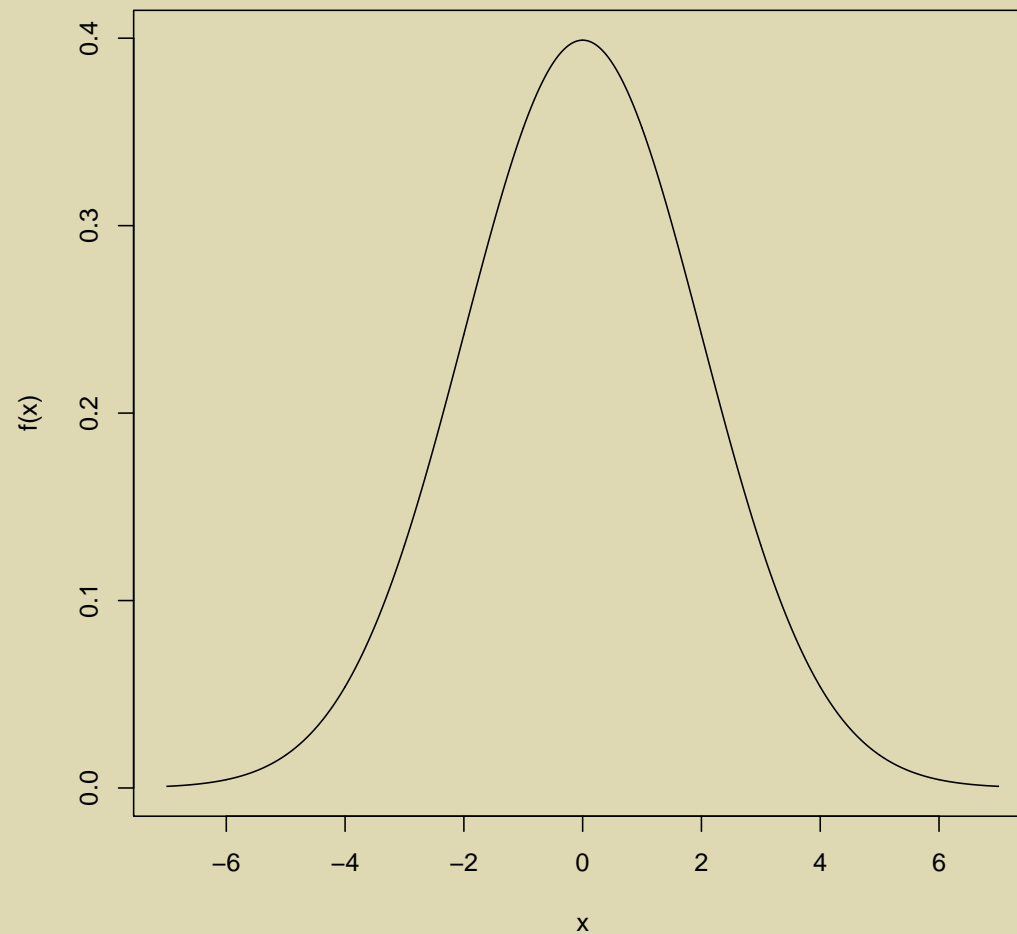
Outliers

Transformations

Wrapping Up

$$N(0, 2)$$

Univariate Normal Distribution



# Univariate Normal Distribution

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

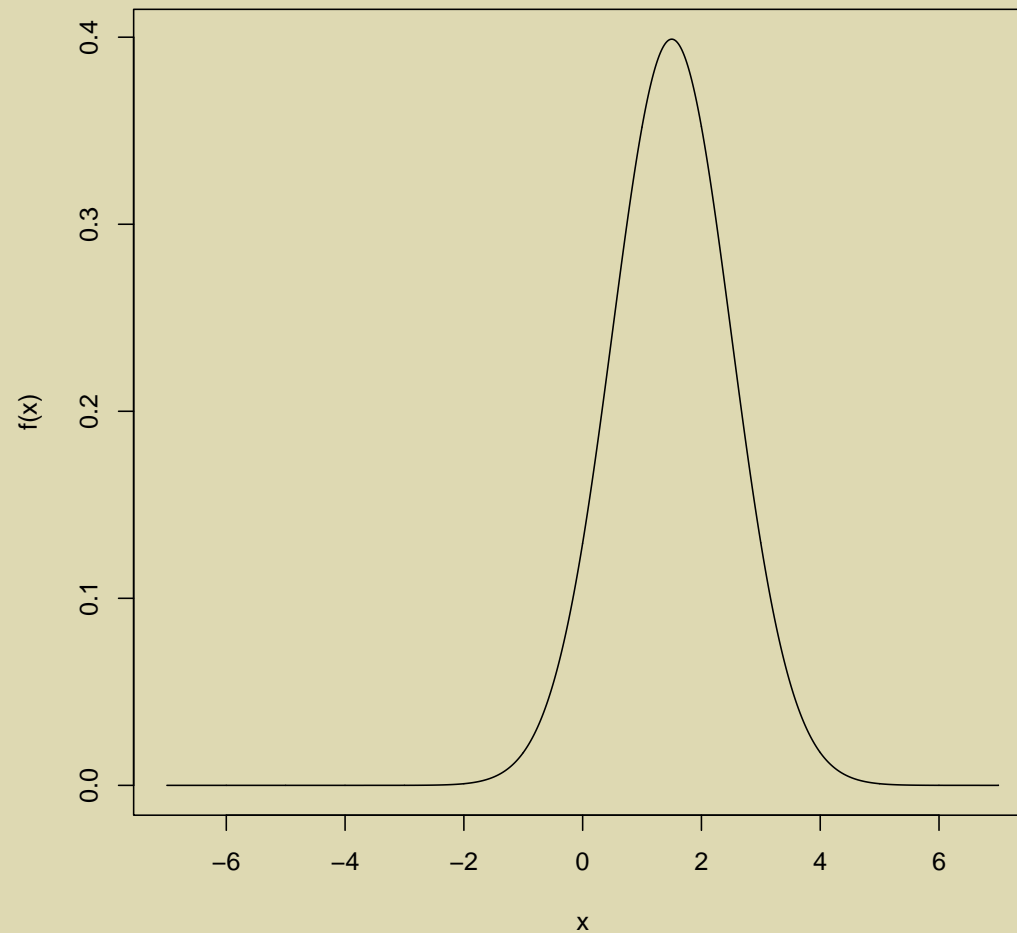
Outliers

Transformations

Wrapping Up

$$N(3, 1)$$

Univariate Normal Distribution



Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Recall that the area under the curve for the univariate normal distribution is a function of the variance/standard deviation.

- In particular:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.954$$

- Also note the term in the exponent:

$$\left(\frac{(x - \mu)}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$$

- This is the square of the distance from  $x$  to  $\mu$  in standard deviation units, and will be generalized for the MVN.

[Overview](#)[Generalized Variance](#)[MVN](#)[Univariate Review](#)[MVN](#)[MVN Contours](#)[MVN Properties](#)[Distributions](#)[Assessing Uni Normality](#)[Assessing MV Normality](#)[Outliers](#)[Transformations](#)[Wrapping Up](#)

- The multivariate normal distribution function is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- The mean vector is  $\boldsymbol{\mu}$ .
- The covariance matrix is  $\Sigma$ .
- Standard notation for multivariate normal distributions is  $N_p(\boldsymbol{\mu}, \Sigma)$ .
- Visualizing the MVN is difficult for more than two dimensions, so I will demonstrate some plots with two variables - the bivariate normal distribution.

# Bivariate Normal Plot #1

Overview

Generalized Variance

MVN

Univariate Review

**MVN**

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

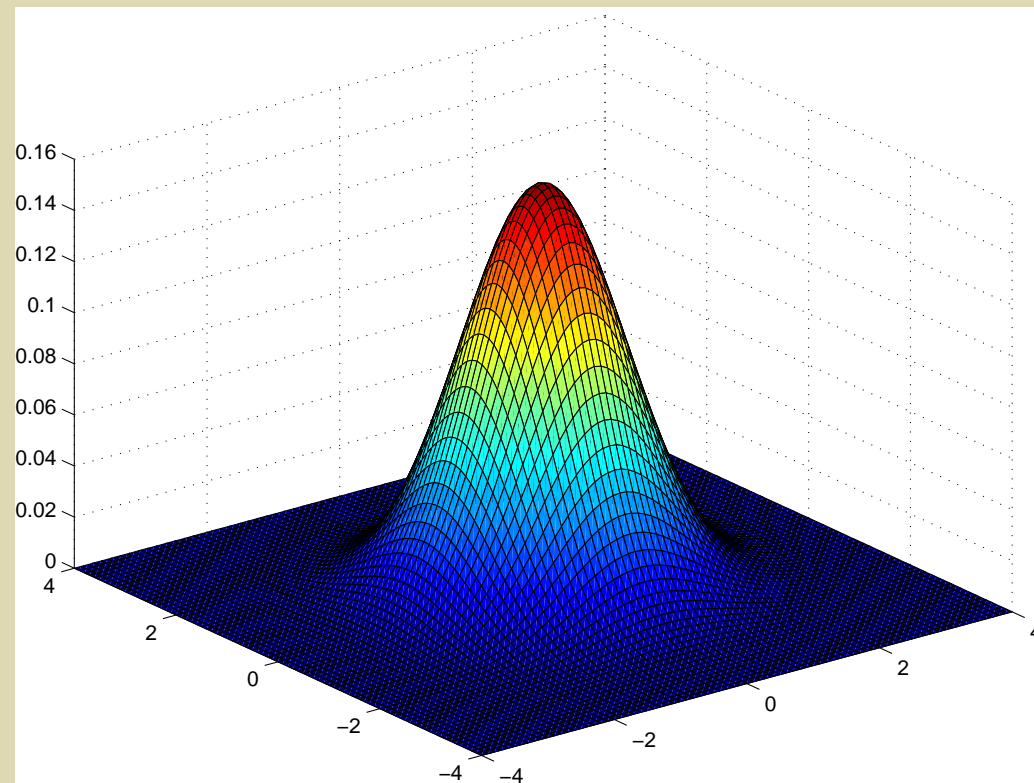
Assessing MV Normality

Outliers

Transformations

Wrapping Up

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



# Bivariate Normal Plot #1a

Overview

Generalized Variance

MVN

Univariate Review

**MVN**

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

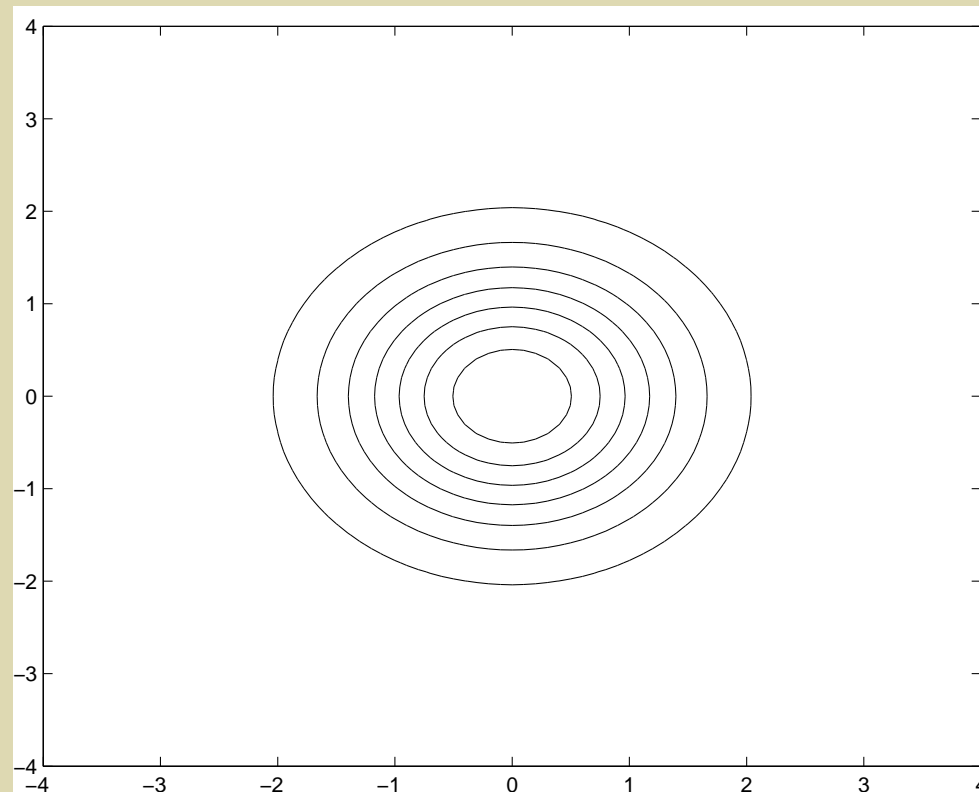
Assessing MV Normality

Outliers

Transformations

Wrapping Up

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



# Bivariate Normal Plot #2

Overview

Generalized Variance

MVN

Univariate Review

**MVN**

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

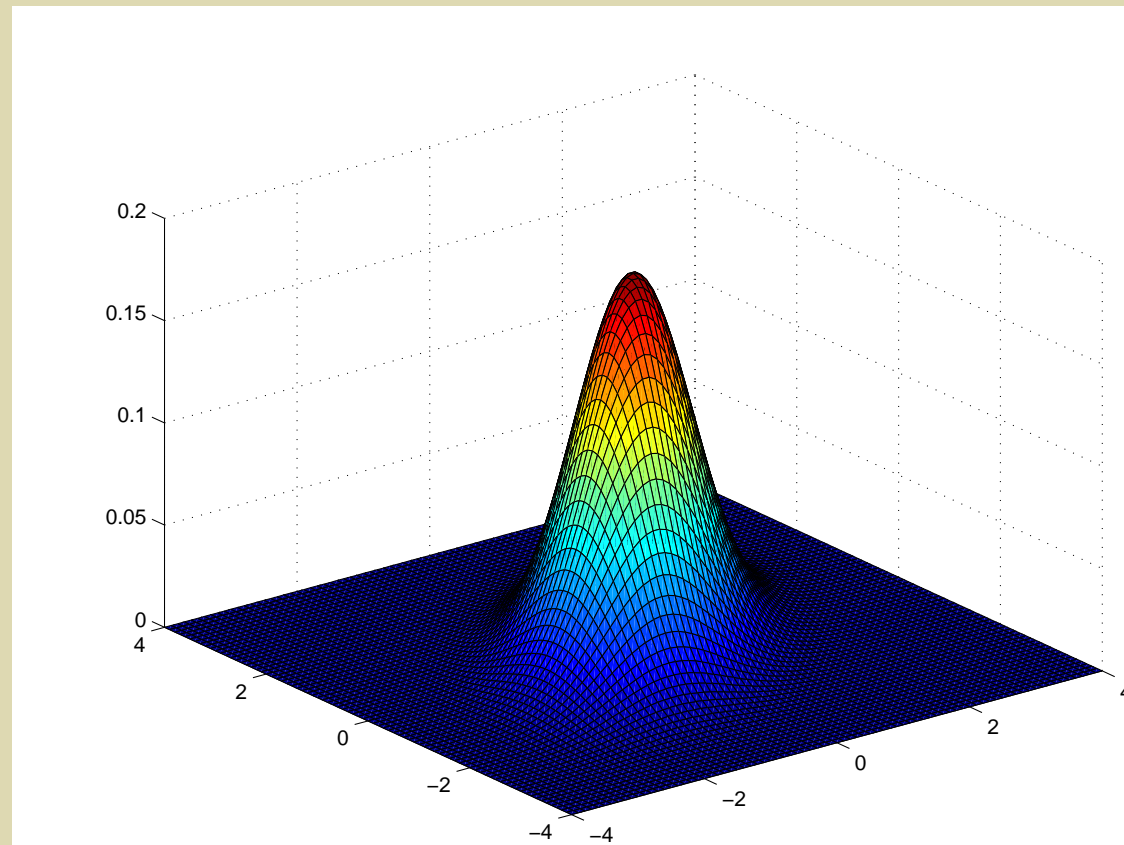
Assessing MV Normality

Outliers

Transformations

Wrapping Up

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



# Bivariate Normal Plot #2

Overview

Generalized Variance

MVN

Univariate Review

**MVN**

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

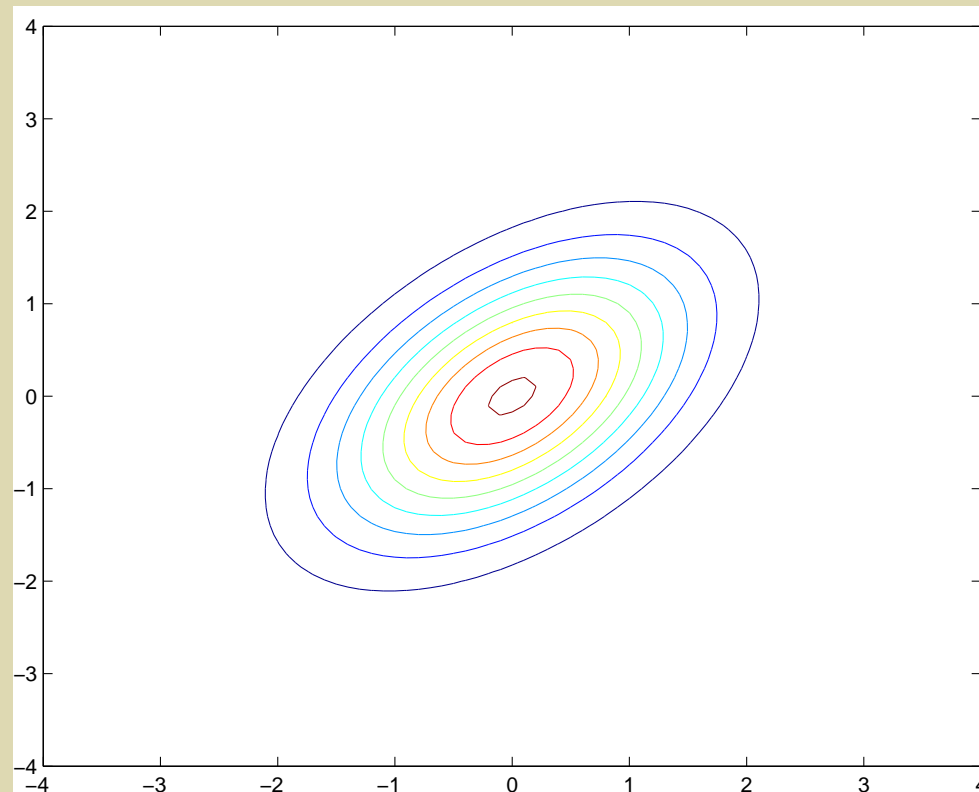
Assessing MV Normality

Outliers

Transformations

Wrapping Up

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



# MVN Contours

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- The lines of the contour plots denote places of equal probability mass for the MVN distribution.
- These contours can be constructed from the eigenvalues and eigenvectors of the covariance matrix.
  - ◆ The direction of the ellipse axes are in the direction of the eigenvalues.
  - ◆ The length of the ellipse axes are proportional to the constant times the eigenvector.
- Specifically:

$$(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$

has ellipsoids centered at  $\boldsymbol{\mu}$ , and has axes  $\pm c\sqrt{\lambda_i}\mathbf{e}_i$ .

# MVN Contours, Continued

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Contours are useful because they provide confidence regions for data points from the MVN distribution.
- The multivariate analog of a confidence interval is given by an ellipsoid, where  $c$  is from the Chi-Squared distribution with  $p$  degrees of freedom.
- Specifically:

$$(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \chi_p^2(\alpha)$$

provides the confidence region containing  $1 - \alpha$  of the probability mass of the MVN distribution.

# MVN Contour Example

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Imagine we had a bivariate normal distribution with:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- The covariance matrix has eigenvalues and eigenvectors:

$$\boldsymbol{\lambda} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}, \mathbf{E} = \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix}$$

- We want to find a contour where 95% of the probability will fall, corresponding to  $\chi_2^2(0.05) = 5.99$

# MVN Contour Example

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- This contour will be centered at  $\mu$ .

- Axis 1:

$$\mu \pm \sqrt{5.99 \times 1.5} \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} 2.12 \\ 2.12 \end{bmatrix}, \begin{bmatrix} -2.12 \\ -2.12 \end{bmatrix}$$

- Axis 2:

$$\mu \pm \sqrt{5.99 \times 0.5} \begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} -1.22 \\ 1.22 \end{bmatrix}, \begin{bmatrix} 1.22 \\ -1.22 \end{bmatrix}$$

# MVN Properties

Overview

Generalized Variance

MVN

Univariate Review

MVN

MVN Contours

MVN Properties

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- The MVN distribution has some convenient properties.
- If  $\mathbf{X}$  has a multivariate normal distribution, then:
  1. Linear combinations of  $\mathbf{X}$  are normally distributed.
  2. All subsets of the components of  $\mathbf{X}$  have a MVN distribution.
  3. Zero covariance implies that the corresponding components are independently distributed.
  4. The conditional distributions of the components are MVN.

# Distribution of $\bar{x}$ and $S$

Overview

Generalized Variance

MVN

Distributions

UVN CLT

Multi CLT

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Recall back in Univariate statistics you discussed the Central Limit Theorem (CLT)

It stated that, if the set of  $n$  observations  $x_1, x_2, \dots, x_n$  were normal or not...

- The distribution of  $\bar{x}$  would be normal with mean equal to  $\mu$  and variance  $\sigma^2/n$
- We were also told that  $(n-1)s^2/\sigma^2$  had a Chi-Square distribution with  $n-1$  degrees of freedom
- *Note: We ended up using these pieces of information for hypothesis testing such as t-test and ANOVA.*

# Distribution of $\bar{\mathbf{x}}$ and $\mathbf{S}$

Overview

Generalized Variance

MVN

Distributions

UVN CLT

Multi CLT

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

We also have a Multivariate Central Limit Theorem (CLT)

It states that, if the set of  $n$  observations  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  are multivariate normal or not...

- The distribution of  $\bar{\mathbf{x}}$  would be normal with mean equal to  $\mu$  and variance/covariance matrix  $\Sigma/n$
- We are also told that  $(n - 1)\mathbf{S}$  will have a Wishart distribution,  $W_p(n - 1, \Sigma)$ , with  $n - 1$  degrees of freedom
  - ◆ This is the multivariate analogue to a Chi-Square distribution.
- *Note: We will end up using some of this information for multivariate hypothesis testing.*

# Distribution of $\bar{\mathbf{X}}$ and $\mathbf{S}$

Overview

Generalized Variance

MVN

Distributions

UVN CLT

Multi CLT

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Therefore, let  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  be independent observations from a population with mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$
- The following are true:
  - ◆  $\sqrt{n} (\bar{\mathbf{X}} - \boldsymbol{\mu})$  is approximately  $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ .
  - ◆  $n (\mathbf{X} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\mathbf{X} - \boldsymbol{\mu})$  is approximately  $\chi_p^2$ .

# Assessing Normality

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- Recall from earlier that IF the data have a Multivariate normal distribution then all of the previously discussed properties will hold.
- So we will want to have at least a set of test/methods to assess Multivariate Normality.
  - ◆ We said that if all marginal distributions are NOT normal then the joint distribution can not be MVN. So we will first talk about the assessing normality
  - ◆ We also said even if all marginals are normally distributed the joint is not necessarily MVN, So we will then assess Multivariate Normality.

# Assessing Normality

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- We will find that there are two ways to assess normality/MVN.
  1. By comparing the distribution of your observations (or some transformation of your observations) to some known distribution. (These are commonly called Q-Q plots)
  2. By computing some set of statistics and obtaining a p-value (i.e., compute a statistic with a known distribution and determine how extreme the statistic is compared to a null hypothesis).

# Assessing Univariate Normality

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

We begin with Assessing Univariate Normality using a Q-Q plot.

- A Q-Q plot is a plot that matches the Quantiles of the observed data with the Quantiles of a specific distribution.
- A Quantile (commonly called a percentile) is that value such that a specific proportion  $p$  of the population will score at or below.
  - ◆ For example the .5 quantile of a  $N(0,1)$  is 0.
- In our case the Quantiles of a specific distribution will be a normal,  $N(0, 1)$ .
  - ◆ It could be a  $N(\bar{x}, s_x^2)$ , if preferred.
- There should be a linear relationship between the quantiles of the observed data with their theoretical quantiles (assuming the distribution) if they follow the same distribution.

# Constructing a Q-Q plot

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Lets assume that we have  $n$  observations  $x_1, x_2, \dots, x_n$ . To construct a Q-Q plot we:

1. Order the observations from smallest to largest (i.e.,  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  ).
2. Next we define the  $i^{th}$  point,  $x_{(i)}$ , as the  $(i - .5)/n$  quantile.
  - We could use  $i/n$  but can cause problems.
3. Based on a  $N(0, 1)$  distribution we compute the quantile values  $q_1, q_2, \dots, q_n$  (this is typically done using a table or computer).
4. Finally plot  $(x_{(i)}, q_i)$ , and if they follow the same distribution (Normal) they should form a line.

# Example Q-Q plot

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Lets assume that we have 5 observations: 3, 6, 4, 5, 2:

First we order them

$y_{(i)}$	$(i - .5)/n$	$q_i$
2		
3		
4		
5		
6		

# Example Q-Q plot

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Lets assume that we have 5 observations: 3, 6, 4, 5, 2:

Next compute quantiles

$y(i)$	$(i - .5)/n$	$q_i$
2	$(1 - .5)/5 = .1$	
3	$(2 - .5)/5 = .3$	
4	$(3 - .5)/5 = .5$	
5	$(4 - .5)/5 = .7$	
6	$(5 - .5)/5 = .9$	

# Example Q-Q plot

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Lets assume that we have 5 observations: 3, 6, 4, 5, 2:

Finally compute quantiles values assuming  $N(0, 1)$  (i.e., this is a z-score)

$y(i)$	$(i - .5)/n$	$q_i$
2	$(1 - .5)/5 = .1$	-1.28
3	$(2 - .5)/5 = .3$	-0.52
4	$(3 - .5)/5 = .5$	0.00
5	$(4 - .5)/5 = .7$	0.52
6	$(5 - .5)/5 = .9$	1.28

and plot

# Example Q-Q plot

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

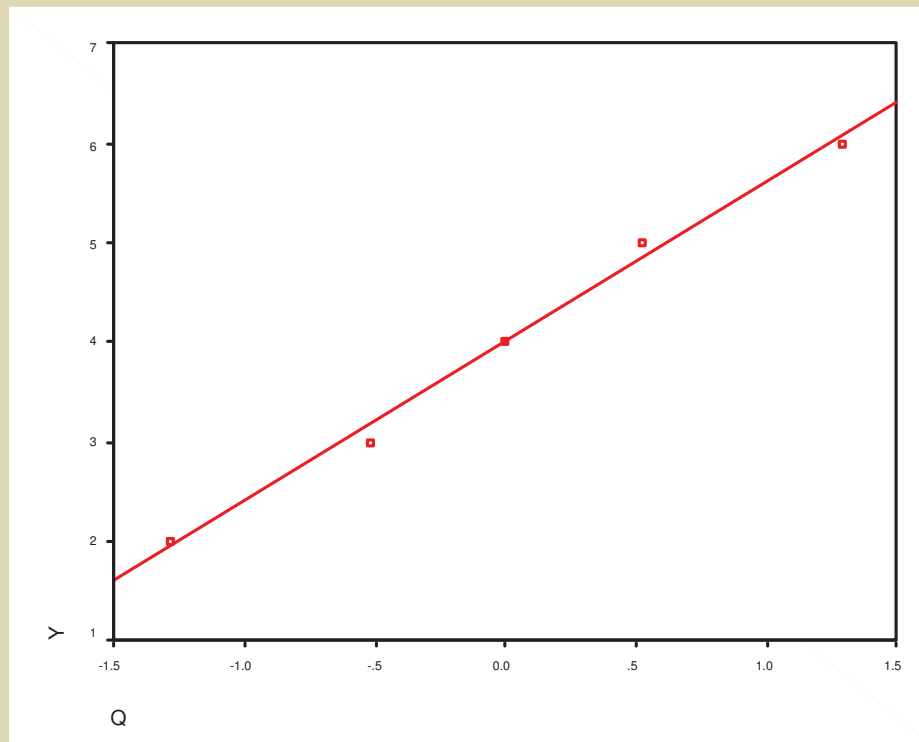
Outliers

Transformations

Wrapping Up

Notice how it follows nearly a straight line

Figure 1: Q-Q plot



# Three Tests for Normality

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

- The remaining methods are tests for normality
- In each case it is computing a statistic and checking for significance.
- I will mention these because in some journals this is mentioned.
- However, because we will be mostly interested in MVN one could quickly check for normality and the if happy, test for MVN.

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Begin by computing Skewness and Kurtosis

Skewness,  $\sqrt{b_1}$

$$\sqrt{b_1} = \frac{\sqrt{n} \sum_{i=1}^n (y_i - \bar{y})^3}{[\sum_{i=1}^n (y_i - \bar{y})^2]^{3/2}}$$

Kurtosis,  $b_2$

$$b_2 = \frac{n \sum_{i=1}^n (y_i - \bar{y})^4}{[\sum_{i=1}^n (y_i - \bar{y})^2]^2}$$

# Other Tests

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing Normality

Uni Norm

Make a Q-Q plot

Example Q-Q plot

Normality Tests

Test 1

Other Tests

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Other tests can be gathered in SAS (note the null hypothesis is always that the data is normally distributed).

```
proc univariate data=mydata normal plot;  
var x1-x5;  
run;
```

# Assessing MVN

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Assessing MVN

Scatter Plots

Q-Q Plots

Tests

Outliers

Transformations

Wrapping Up

- Notice that many of the procedures that we discussed (Including the Q-Q plot) required that we rank order the observations.
- If instead of single observations, we have a set of  $n$  observations of  $p$  variables ( $\mathbf{x}_1, \mathbf{x}_1, \dots, \mathbf{x}_n$ ) ordering all observations is much more difficult.
- In fact, the book mentions that any test for MVN is far more difficult and often has little power because of the number of observations on  $p$ -space, but some check should be done.

# Scatter Plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Assessing MVN

Scatter Plots

Q-Q Plots

Tests

Outliers

Transformations

Wrapping Up

Here we will discuss two graphical methods

- Possibly the easiest method to assess MVN is to use scatter plots.
- If there are not a large number of variables consider looking scatter plots of all pairs of variables.
  - ◆ Recall that one property of a MVN is that all subsets of variables should also be multivariate (in this case bivariate) normal.
  - ◆ This means any relationship between variables should be linear and otherwise should have a random pattern
- Could also look at all sets of three variables should also have tri-variate normal distributions.

# Q-Q Plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Assessing MVN

Scatter Plots

Q-Q Plots

Tests

Outliers

Transformations

Wrapping Up

As an alternative (for example if there are too many variables we could use a Q-Q plot

- Our Q-Q plot will be based on  $D^2 = (\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}})$ . Note,  $\mathbf{x}$  is each entity.

- ◆ Recall that  $D^2 = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$  has a  $\chi_p^2$  distribution with  $p$  degrees of freedom.
- ◆ We could use that to compute a Q-Q plot (i.e., use the  $\chi_p^2$  distribution to get the values of  $q_i$ ).
- ◆ However in estimating the mean vector and variance covariance matrix this plot could be misleading.

- Instead, some suggest using a function of  $D^2$ .

# Q-Q Plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Assessing MVN

Scatter Plots

Q-Q Plots

Tests

Outliers

Transformations

Wrapping Up

We will define a new variable  $u_i$ , where

$$u_i = D_i^2.$$

1. Order the observations from smallest to largest (i.e.,  $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$  ).
2. Next we define the  $i^{th}$  point,  $u_{(i)}$ , as the  $(i - .5)/n$  quantile.
3. Based on a  $\chi_p^2$  distribution, we compute the quantile values  $q_1, q_2, \dots, q_n$  (this is typically done using a table or computer) where:

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Assessing MVN

Scatter Plots

Q-Q Plots

Tests

Outliers

Transformations

Wrapping Up

- Tests based on the Skewness and Kurtosis also exist.
- These are a generalizations of the univariate tests.
- You will not be tested on this, but know where it is just in case you ever need a test.

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

- Detecting outliers is difficult in Multiple dimensions.
  - ◆ Can't simply order them and look for extreme values.
  - ◆ May not be able to see in only 2-D plots.
  - ◆ Could be different degrees of recording errors.
- Here I discuss one method for detecting Outliers

# Outliers Using $D^2$

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

Wilk's statistics was designed to detect a single outlier:

- Compute  $w$  where

$$w = \max_i \frac{|(n-2)\mathbf{S}_{-i}|}{|(n-1)\mathbf{S}|}$$

- To simplify it can be shown that  $w$  also equals:

$$w = 1 - \frac{nD_{(n)}^2}{(n-1)^2}$$

- The distribution for  $w$  is the  $F$  distribution, however the only unknown is  $D_{(n)}^2$  and so we can just make decisions based on it.
- Therefore a test can be made that is based on the  $D^2$ , which is also computed to assess MVN

# Example

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

Here we consider three variables: time it takes people to get to class, number of years in school, overall happiness.

- Just as a generic procedure I would
  1. Evaluate each variable individually
  2. Evaluate each pair (if this is reasonable)
  3. Evaluate MVN using a Q-Q plot
  4. Evaluate the reasonableness of any outliers (this can also be done in the previous steps).

# Example Q-Q plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

Example Bivariate plots

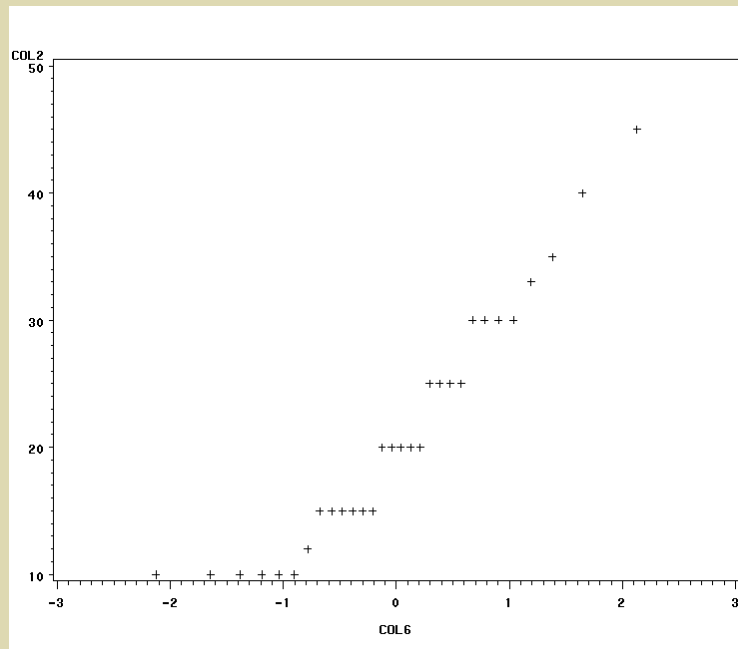
MV Q-Q Plot

Transformations

Wrapping Up

First we use a Q-Q plot for each variable.

Figure 2: Q-Q plot of Time



# Example Q-Q plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

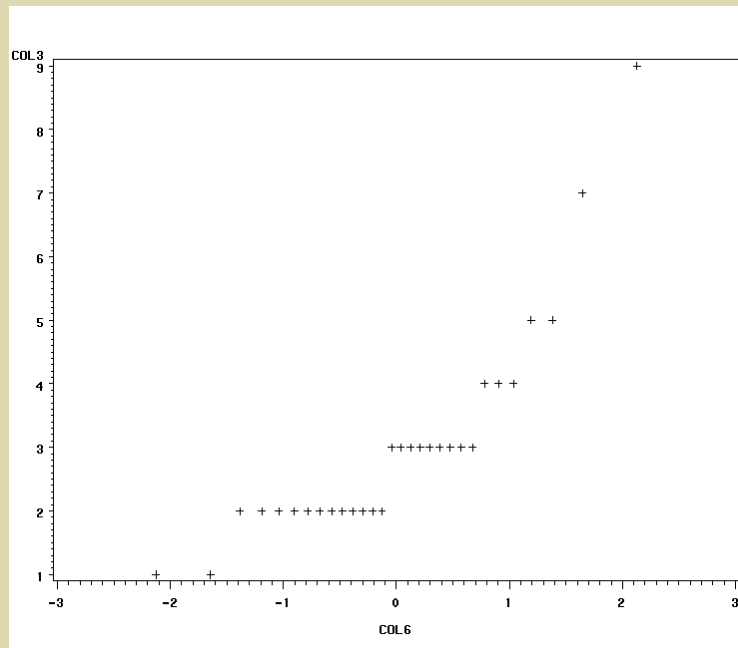
Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

Figure 3: Q-Q plot of Year



# Example Q-Q plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

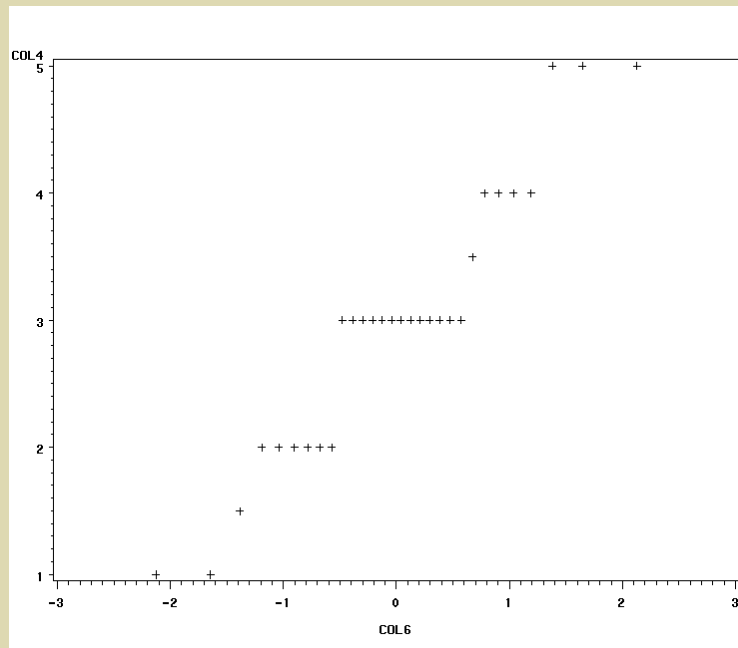
Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

Figure 4: Q-Q plot of Happy



# Example Bivariate plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

- So if we decide that everything looks OK then we move on to the bivariate plots.
- If they do not look OK we can consider using some kind of robust analyses or consider a transformation.

# Example Bivariate plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

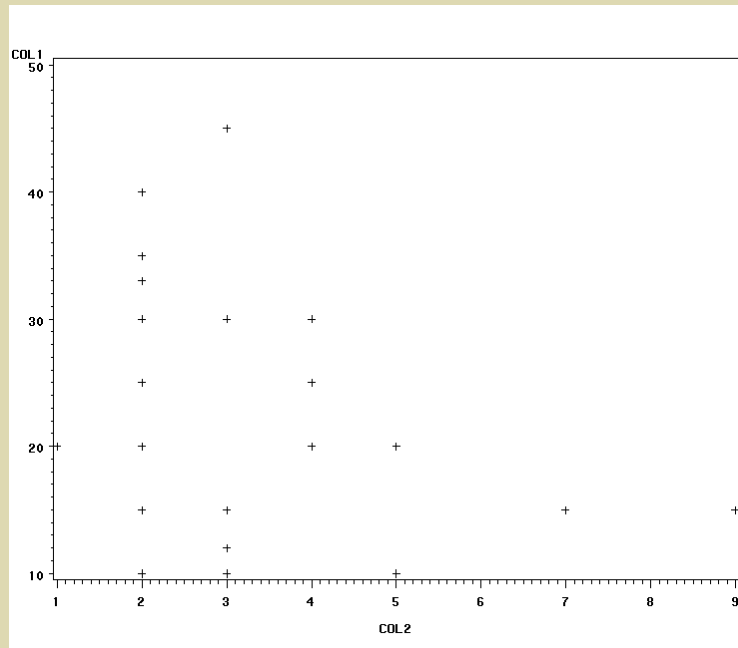
Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

Figure 5: Time versus Year



# Example Bivariate plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

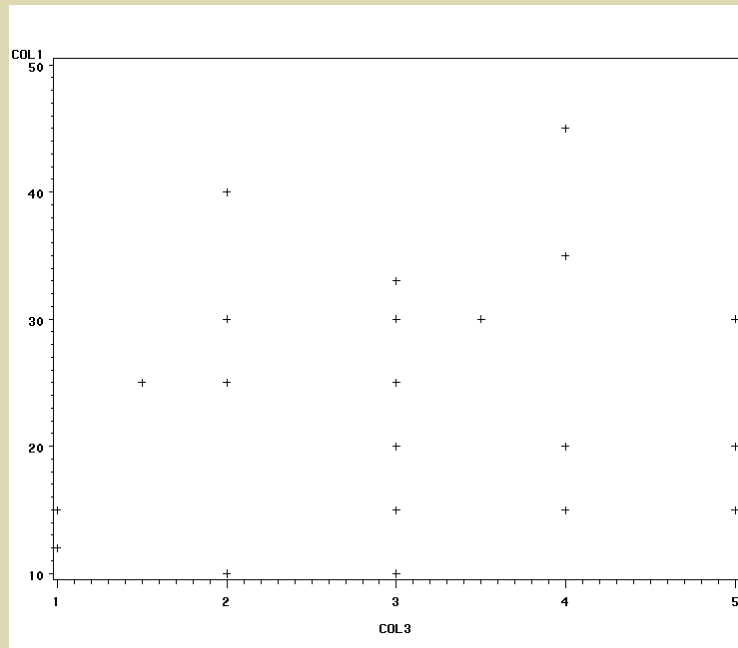
Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

Figure 6: Time versus Happy



# Example Bivariate plots

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

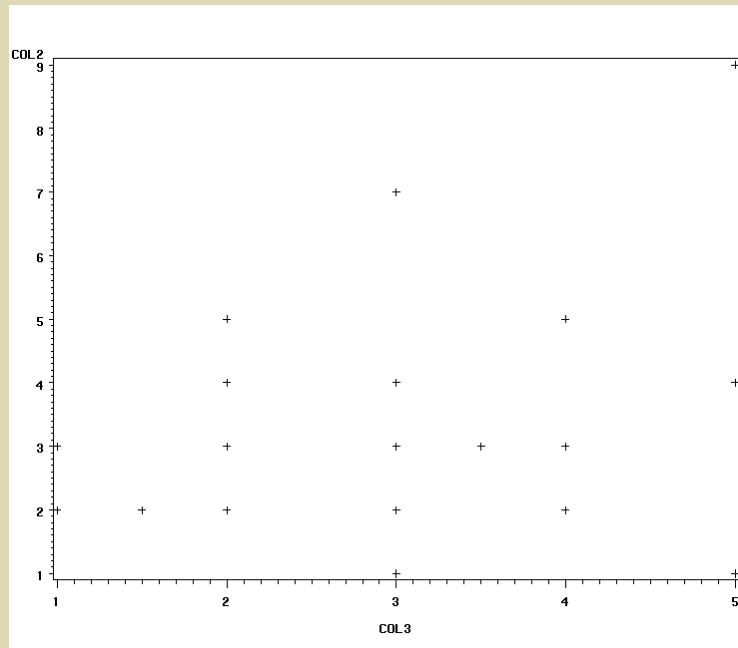
Example Bivariate plots

MV Q-Q Plot

Transformations

Wrapping Up

Figure 7: Year versus Happy



# MV Q-Q Plot

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Outliers

Outliers Using  $D^2$

Example

Example Q-Q plots

Example Bivariate plots

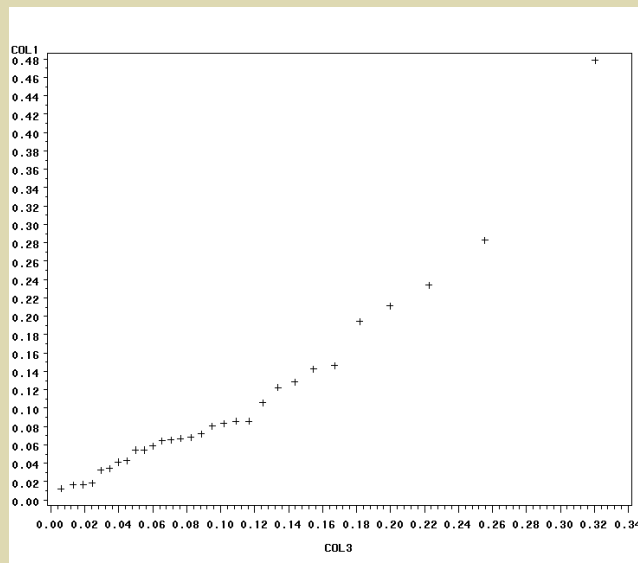
MV Q-Q Plot

Transformations

Wrapping Up

Finally, the Multivariate Q-Q plot.

Figure 8: Year versus Happy



- We can also look for an outlier.
- Largest  $D^2 = 13.4$
- Compare it to Table A.6 for 30 observation and 3 variables.
  - ◆ Max  $D^2$  is 12.24 with p-value=.05 and 14.14 with p-value=.01.

# Transformations to Near Normality

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Transformations to Near  
Normality

Wrapping Up

A few types of data must be transformed prior to doing analyses that assume data comes from a MVN distribution:

- Counts,  $y$ , are transformed with  $\sqrt{y}$ .
- Proportions,  $\hat{p}$ , are transformed with  $\text{logit}(\hat{p}) = \frac{1}{2} \log \left( \frac{\hat{p}}{1-\hat{p}} \right)$
- Correlations,  $r$ , are transformed with (Fisher's)  
$$z(r) = \frac{1}{2} \log \left( \frac{1+r}{1-r} \right).$$

Variations of transformations exist, as do other techniques.

# Final Thought

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Final Thought

Next Class

- The multivariate normal distribution is an analog to the univariate normal distribution.
- The MVN distribution will play a large role in the upcoming weeks.
- We can finally put the background material to rest, and begin learning some practical statistics.



# Next Time

Overview

Generalized Variance

MVN

Distributions

Assessing Uni Normality

Assessing MV Normality

Outliers

Transformations

Wrapping Up

Final Thought

Next Class

## ■ Statistical Analyses - Mean Vector Inference.