# Multivariate Statistical Analysis

Shyh-Kang Jeng
Department of Electrical Engineering/
Graduate Institute of Communication/
Graduate Institute of Networking and Multimedia

1

## Outline

- Introduction
- Organization of Data
- Data Displays and Pictorial Representations
- Distances
- Reading Assignments

2

## Outline

- Introduction
- Organization of Data
- Data Displays and Pictorial Representations
- Distances
- Reading Assignments

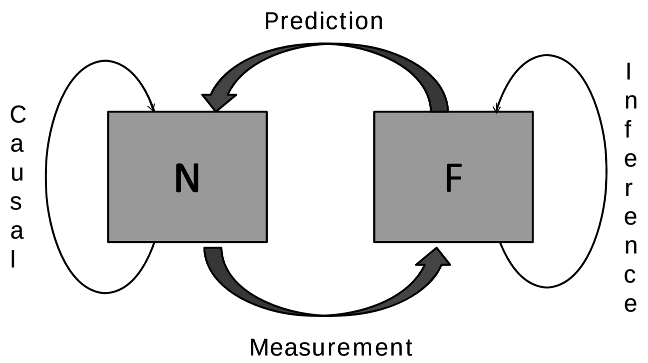
3

## Questions

- What is a model?
- How to model Nature?
- What is statistics?

4

# How to Model Nature?




R. Rosen, Life Itself, Columbia Univ. Press, 1991

5

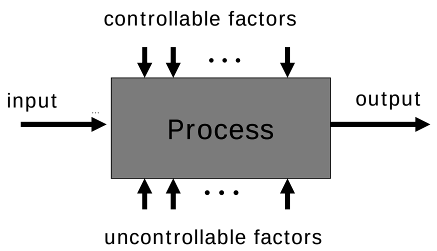
# Questions

- What is univariate statistics?
- What is multivariate statistics?
- Why to learn multivariate analysis?
- What are major uses and applications of multivariate analysis?
- What will be covered in this course?
- What are required to this course?
- How to handle the term project?

6

# What Is Multivariate Analysis?

- Statistical methodology to analyze data with measurements on many variables




7

# Why to Learn Multivariate Analysis?

- Explanation of a social or physical phenomenon must be tested by gathering and analyzing data
- Complexities of most phenomena require an investigator to collect observations on many different variables

8

## Major Uses of Multivariate Analysis

- Data reduction or structural simplification
- Sorting and grouping
- Investigation of the dependence among variables
- Prediction
- Hypothesis construction and testing

9

## Application Examples

- Is one product better than the other?
- Which factor is the most important to determine the performance of a system?
- How to classify the results into clusters?
- What are the relationships between variables?

10

## Course Outline

- Introduction
- Matrix Algebra and Random Vectors
- Sample Geometry and Random Samples
- Multivariate Normal Distribution
- Inference about a Mean Vector
- Comparison of Several Multivariate Means
- Multivariate Linear Regression Models

11

## Course Outline

- Principal Components
- Factor Analysis and Inference for Structured Covariance Matrices
- Canonical Correlation Analysis*
- Discrimination and Classification*
- Clustering, Distance Methods, and Ordination*

12

## Important Multivariate Techniques Not Included

- Structural Equation Models
- Multidimensional Scaling

13

## Feature of This Course

- Uses matrix algebra to introduce theories and practices of multivariate statistical analysis

14

## Text Book and Website

- R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis, 6th ed., Pearson Education, 2007. (雙葉)
- http://cc.ee.ntu.edu.tw/~skjeng/MultivariateAnalysis2011.htm

15

## References


- 林震岩, 多變量分析-SPSS的操作與應用, 智勝, 2007
- J. F. Hair, Jr., B. Black, B. Babin, R. E. Anderson, and R. L. Tatham, Multivariate Data Analysis, 6th ed., Prentice Hall, 2006. (華泰)
- D. C. Montgomery, Design and Analysis of Experiments, 6th ed., John Wiley, 2005. (歐亞)

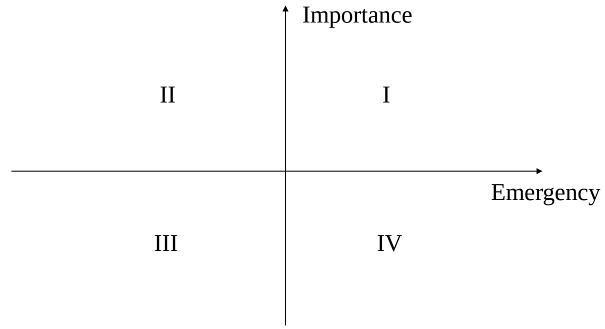

16

## References


- D. Salsberg著, 葉偉文譯,*統計,改變了世界*, 天下遠見, 2001.
- 張碧波，*推理統計學*，三民，1976.
- 張輝煌編譯，*實驗設計與變異分析*，建興，1986.




## Time Management






## Some Important Laws

- First things first
- 80 – 20 Law
- Fast prototyping and evolution
- 物有本末，事有始終，知所先後，則近道矣。



## Outline

- Introduction
- Organization of Data
- Data Displays and Pictorial Representations
- Distances
- Reading Assignments

## Questions

- How to represent the measurement data for multivariate analysis?
- How to summarize the measurement data?
- How to determine if two variables are related?



## Array of Data

$$\mathbf{x} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$



## Descriptive Statistics

- Summary numbers to assess the information contained in data
- Basic descriptive statistics
  - Sample mean
  - Sample variance
  - Sample standard deviation
  - Sample covariance
  - Sample correlation coefficient



## Sample Mean and Sample Variance

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$$

$$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2$$

$$k = 1, 2, \cdots, p$$

## Sample Covariance and Sample Correlation Coefficient

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji}-\bar{x}_i\right)\left(x_{jk}-\bar{x}_k\right)$$

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}\left(x_{ji}-\bar{x}_i\right)\left(x_{jk}-\bar{x}_k\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{ji}-\bar{x}_i\right)^2}\sqrt{\sum_{j=1}^{n}\left(x_{jk}-\bar{x}_k\right)^2}}$$

$$i = 1, 2, \cdots, p; \quad k = 1, 2, \cdots, p$$

$$s_{ik} = s_{ki}, \quad r_{ik} = r_{ki}$$

25

## Standardized Values (or Standardized Scores)

- Centered at zero
- Unit standard deviation
- Sample correlation coefficient can be regarded as a sample covariance of two standardized variables

$$\frac{x_{jk}-\bar{x}_k}{\sqrt{s_{kk}}}$$

26

## Properties of Sample Correlation Coefficient

- Value is between -1 and 1
- Magnitude measure the strength of the linear association
- Sign indicates the direction of the association
- Value remains unchanged if all $x_{ji}$'s and $x_{jk}$'s are changed to $y_{ji} = a\, x_{ji} + b$ and $y_{jk} = c\, x_{jk} + d$, respectively, provided that the constants $a$ and $c$ have the same sign

27

## Arrays of Basic Descriptive Statistics

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}, \quad \mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

28

## Example

- Four receipts from a university bookstore
- Variable 1: dollar sales
- Variable 2: number of books

$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

29

## Arrays of Basic Descriptive Statistics

$$\bar{\mathbf{x}} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}, \quad \mathbf{S}_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

30

## Outline

- Introduction
- Organization of Data
- <u>Data Displays and Pictorial Representations</u>
- Distances
- Reading Assignments

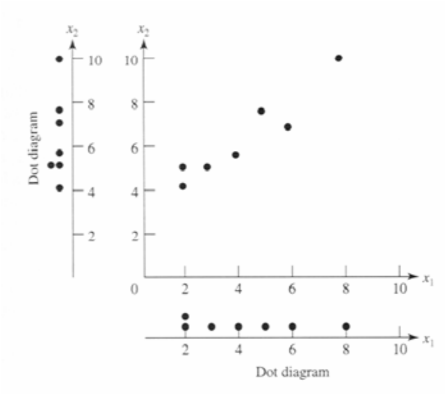
31

## Questions

- How to visually represent multivariate data?
- What are the advantages of data plots?

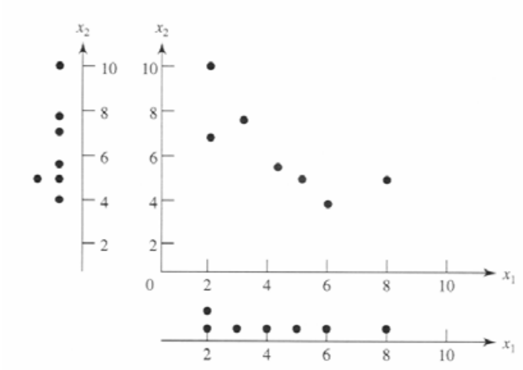
32

## Scatter Plot and Marginal Dot Diagrams



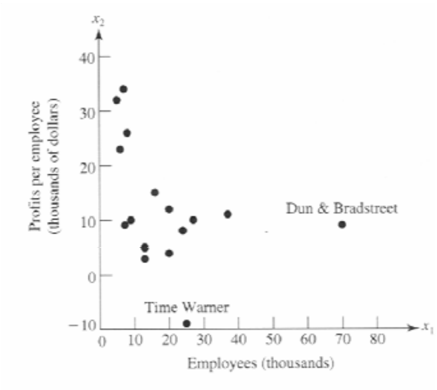



## Scatter Plot and Marginal Dot Diagrams for Rearranged Data






## Effect of Unusual Observations





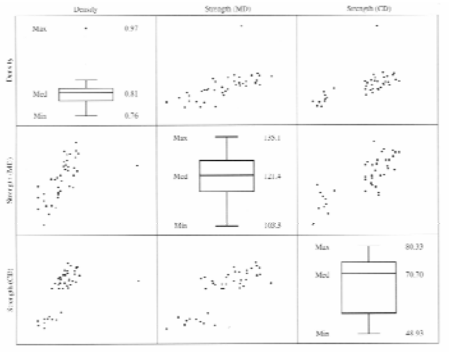

## Effect of Unusual Observations

$$r_{12} = \begin{cases} -0.39 & \text{for all 16 firms} \\ -0.56 & \text{for all firms but Dun \& Bradstreet} \\ -0.39 & \text{for all firms but Time Warner} \\ -0.50 & \text{for all firms but Dun \& Bradstreet and Time Warner} \end{cases}$$

## Paper Quality Measurements

| Specimen | Density | Strength: Machine direction | Strength: Cross direction |
|---|---|---|---|
| 1 | .801 | 121.41 | 70.42 |
| 2 | .824 | 127.70 | 72.47 |
| 3 | .841 | 129.20 | 78.20 |
| 4 | .816 | 131.80 | 74.89 |
| 5 | .840 | 135.10 | 71.21 |
| 6 | .842 | 131.50 | 78.39 |
| 7 | .820 | 126.70 | 69.02 |
| 8 | .802 | 115.10 | 73.10 |
| 9 | .828 | 130.80 | 79.28 |
| 10 | .819 | 124.60 | 76.48 |
| 11 | .826 | 118.31 | 70.25 |
| 12 | .802 | 114.20 | 72.88 |
| 13 | .810 | 120.30 | 68.23 |
| 14 | .802 | 115.70 | 68.12 |
| 15 | .832 | 117.51 | 71.62 |
| 16 | .796 | 109.81 | 53.10 |
| 17 | .759 | 109.10 | 50.85 |
| 18 | .770 | 115.10 | 51.68 |
| 19 | .759 | 118.31 | 50.60 |
| 20 | .772 | 112.60 | 53.51 |
| 21 | .806 | 116.20 | 56.53 |
| 22 | .803 | 118.00 | 70.70 |
| 23 | .845 | 131.00 | 74.35 |
| 24 | .822 | 125.70 | 68.29 |
| 25 | .971 | 126.10 | 72.10 |
| 26 | .816 | 125.80 | 70.64 |
| 27 | .836 | 125.50 | 76.33 |
| 28 | .815 | 127.80 | 76.75 |
| 29 | .822 | 130.50 | 80.33 |
| 30 | .822 | 127.90 | 75.68 |
| 31 | .843 | 123.90 | 78.54 |
| 32 | .824 | 124.10 | 71.91 |
| 33 | .788 | 120.80 | 68.22 |
| 34 | .782 | 107.40 | 54.42 |
| 35 | .795 | 120.70 | 70.41 |
| 36 | .805 | 121.91 | 73.68 |
| 37 | .836 | 122.31 | 74.93 |
| 38 | .788 | 110.60 | 53.52 |
| 39 | .772 | 103.51 | 48.93 |
| 40 | .776 | 110.71 | 53.67 |
| 41 | .758 | 113.80 | 52.42 |

Source: Data courtesy of SONOCO Products Company.




37

## Lizard Size Data

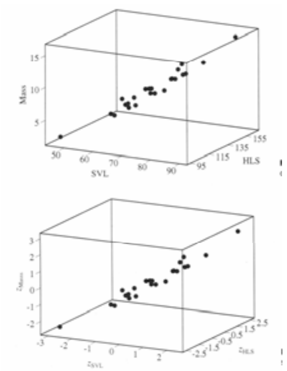
| Lizard | Mass | SVL | HLS | Lizard | Mass | SVL | HLS |
|---|---|---|---|---|---|---|---|
| 1 | 5.526 | 59.0 | 113.5 | 14 | 10.067 | 73.0 | 136.5 |
| 2 | 10.401 | 75.0 | 142.0 | 15 | 10.091 | 73.0 | 135.5 |
| 3 | 9.213 | 69.0 | 124.0 | 16 | 10.888 | 77.0 | 139.0 |
| 4 | 8.953 | 67.5 | 125.0 | 17 | 7.610 | 61.5 | 118.0 |
| 5 | 7.063 | 62.0 | 129.5 | 18 | 7.733 | 66.5 | 133.5 |
| 6 | 6.610 | 62.0 | 123.0 | 19 | 12.015 | 79.5 | 150.0 |
| 7 | 11.273 | 74.0 | 140.0 | 20 | 10.049 | 74.0 | 137.0 |
| 8 | 2.447 | 47.0 | 97.0 | 21 | 5.149 | 59.5 | 116.0 |
| 9 | 15.493 | 86.5 | 162.0 | 22 | 9.158 | 68.0 | 123.0 |
| 10 | 9.004 | 69.0 | 126.5 | 23 | 12.132 | 75.0 | 141.0 |
| 11 | 8.199 | 70.5 | 136.0 | 24 | 6.978 | 66.5 | 117.0 |
| 12 | 6.601 | 64.5 | 116.0 | 25 | 6.890 | 63.0 | 117.0 |
| 13 | 7.622 | 67.5 | 135.0 | | | | |

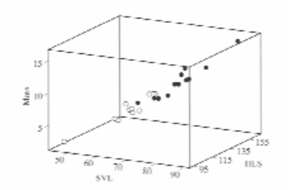
Source: Data courtesy of Kevin E. Bonine.

***SVL: snout-vent length; HLS: hind limb span**

38

## 3D Scatter Plots of Lizard Data



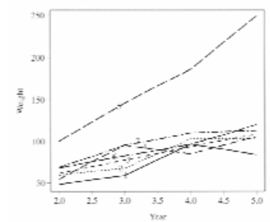




39

## Female Bear Data and Growth Curves

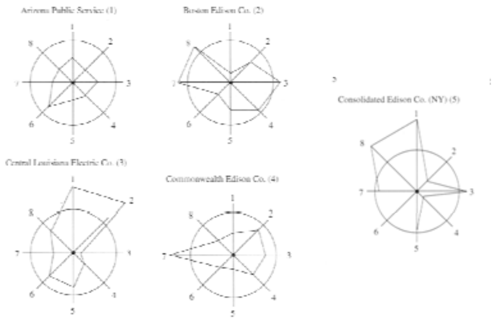
| Bear | Wt2 | Wt3 | Wt4 | Wt5 | Lngth 2 | Lngth 3 | Lngth 4 | Lngth 5 |
|---|---|---|---|---|---|---|---|---|
| 1 | 48 | 59 | 95 | 82 | 141 | 157 | 168 | 183 |
| 2 | 59 | 68 | 102 | 102 | 140 | 168 | 174 | 170 |
| 3 | 61 | 77 | 93 | 107 | 145 | 162 | 172 | 177 |
| 4 | 54 | 43 | 104 | 104 | 146 | 159 | 176 | 171 |
| 5 | 100 | 145 | 185 | 247 | 150 | 158 | 168 | 175 |
| 6 | 68 | 82 | 95 | 118 | 142 | 140 | 178 | 189 |
| 7 | 68 | 95 | 109 | 111 | 139 | 171 | 176 | 175 |

Source: Data courtesy of H. Roberts.



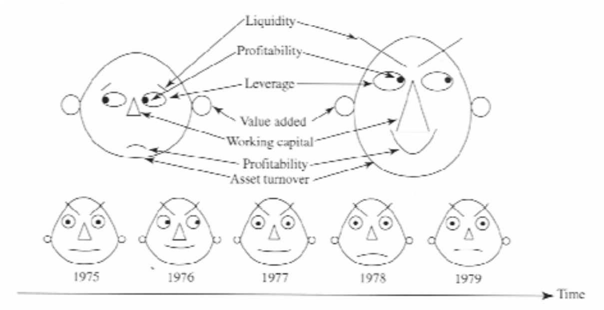

40

## Utility Data as Stars





## Chernoff Faces over Time






## Outline

- Introduction
- Organization of Data
- Data Displays and Pictorial Representations
- Distances
- Reading Assignments



## Questions

- How to determine if two multivariate data are close?
- How to deal with the case that two variables are correlated?

## Euclidean Distance

- Each coordinate contributes equally to the distance

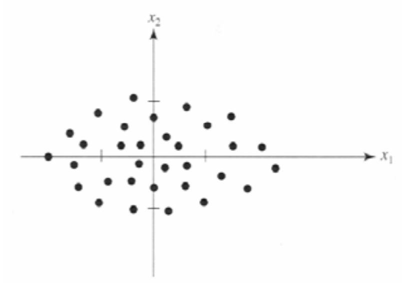
$$P(x_1, x_2, \cdots, x_p), \quad Q(y_1, y_2, \cdots, y_p)$$

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$



## Statistical Distance

- Weight coordinates subject to a great deal of variability less heavily than those that are not highly variable





## Statistical Distance for Uncorrelated Data

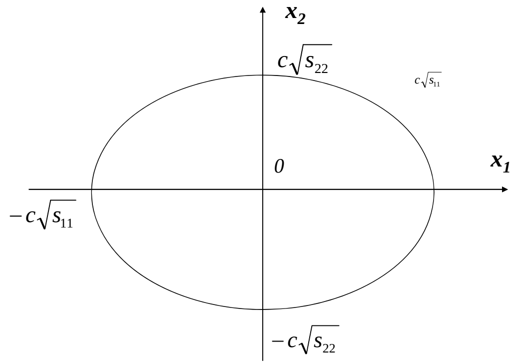
$$P(x_1, x_2), \quad O(0,0)$$

$$x_1^* = x_1 / \sqrt{s_{11}}, \quad x_2^* = x_2 / \sqrt{s_{22}}$$

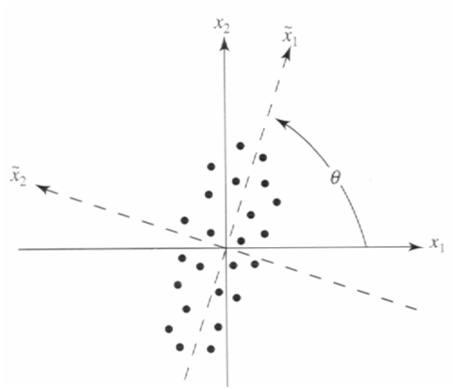
$$d(O,P) = \sqrt{\left(x_1^*\right)^2 + \left(x_2^*\right)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$



## Ellipse of Constant Statistical Distance for Uncorrelated Data



## Scattered Plot for Correlated Measurements




49

## Statistical Distance under Rotated Coordinate System

$$O(0,0), \quad P(\tilde{x}_1, \tilde{x}_2)$$

$$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}}$$

$$\tilde{x}_1 = x_1 \cos\theta + x_2 \sin\theta$$

$$\tilde{x}_2 = -x_1 \sin\theta + x_2 \cos\theta$$

$$d(O,P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

50

## General Statistical Distance

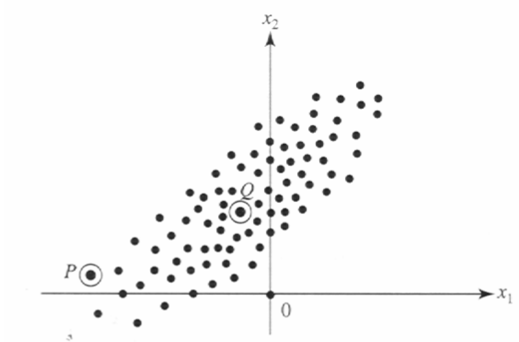
$$P(x_1, x_2, \cdots, x_p), \quad O(0,0,\cdots,0), \quad Q(y_1, y_2, \cdots, y_p)$$

$$d(O,P) = \sqrt{\begin{aligned}&[a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{pp}x_p^2 + \\ &2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{p-1,p}x_{p-1}x_p]\end{aligned}}$$

$$d(P,Q) = \sqrt{\begin{aligned}&[a_{11}(x_1-y_1)^2 + a_{22}(x_2-y_2)^2 + \cdots + \\ &a_{pp}(x_p-y_p)^2 + \\ &2a_{12}(x_1-y_1)(x_2-y_2) + 2a_{13}(x_1-y_1)(x_3-y_3) \\ &+ \cdots + 2a_{p-1,p}(x_{p-1}-y_{p-1})(x_p-y_p)]\end{aligned}}$$

51

## Necessity of Statistical Distance




52

## Necessary Conditions for Statistical Distance Definitions

$$d(P,Q) = d(Q,P)$$

$$d(P,Q) > 0 \text{ if } P \neq Q$$

$$d(P,Q) = 0 \text{ if } P = Q$$

$$d(P,Q) \leq d(P,R) + d(R,Q)$$

$$\text{(Triangle inequality)}$$

53

## Outline

- Introduction
- Organization of Data
- Data Displays and Pictorial Representations
- Distances
- Reading Assignments

54

## Reading Assignments

- Text book
  - pp. 49-59 (Sections 2.1~2.2)
  - pp. 82-96 (Supplement 2A)

55