Statistics 135

Chapter 1 - Multivariate Data

Chris Drake
Department of Statistics
University of California, Davis

Data Arrays

Simultaneous measurements on observational units; relationship between these variables (dependencies, correlations) distinguish univariate from multivariate analysis

Notation: Observational units $u_1, ..., u_n$ with x_{jk} the k^{th} measurement for $k = 1, .., p$ on the j^{th} unit displayed in an array

$$
\begin{array}{ccccccc}
 & x_1 & x_2 & \dots & x_k & \dots & x_p \\
u_1 & x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\
u_2 & x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\
. & . & . & \dots & . & \dots & . \\
u_n & x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np}
\end{array}
$$

Data Summaries

Sample mean $\quad \bar{\mathbf{x}}' = (\bar{x}_1, ..., \bar{x}_p)$

Sample variances and covariances

$$S_n = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

where

$$s_{ii} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2 \qquad s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i) \times (x_{jk} - \bar{x}_k)$$

and

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

with

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii} \times s_{kk}}}$$

Graphical displays

1. Scatterplots (and marginal dot plots)
2. Multiple scatterplots
3. Growth curves for repeated measurements

Distance

Points $P = (x_1, x_2, ..., x_p)$ and $Q = (y_1, y_2, ..., y_p)$

1 Euclidean distance

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

2 Equation of ellipsoid

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2 = c^2$$

3 Statistical distance between two points for uncorrelated variables

$$d(P, Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + ... + \frac{(x_p - y_p)^2}{s_{PP}}}$$

4 Statistical distance for correlated data (2 dimensions)

$$d(P, Q) = \sqrt{a_{11}(x_1 - y_1)^2 + a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$

Note: the coefficients a_{ik} depend on the variability of x_i and x_k and the rotation angle.

5 The equation for an ellipse with center at $(y_1, y_2, ..., y_p)$ is given by

$$c^2 = a_{11}(x_1 - y_1)^2 + a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2$$