

STAT 530/J530 September 1st, 2005

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu



Today

- Multivariate Data
 - * Data Matrix
 - * Summary Parameters and Statistics
 - * A Hint of Matrices
 - * What if Some Data is Missing?
- The Multivariate Normal Distribution



Data!

Multivariate Data is typically presented in a matrix, with each of the n rows being a different observation, and each of the q columns being a different variable.


$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1q} \\ x_{21} & x_{22} & & & & x_{2q} \\ \vdots & & \ddots & & & \vdots \\ x_{i1} & & & x_{ij} & & x_{iq} \\ \vdots & & & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nq} \end{bmatrix}$$



Summarizing Data

Like in univariate statistics the Mean and Variance are commonly used as summary measures.

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \end{bmatrix} \quad \sigma^2 = \begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_q^2 \end{bmatrix}$$


STAT 530/J530 B.Habing Univ. of S.C.  4

Covariances

It is also necessary to summarize the relationship between the variables, and this can be done with the covariance.

$$Cov(x_j, x_k) = E[(x_j - \mu_j)(x_k - \mu_k)]$$


$$\hat{Cov}(x_j, x_k) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

STAT 530/J530 B.Habing Univ. of S.C.  5

Covariance Matrix

The covariance is thus a matrix:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & \sigma_{11} & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{q1} & \cdots & \cdots & \sigma_{qq} \end{bmatrix}$$

STAT 530/J530 B.Habing Univ. of S.C.  6

Sample Covariance Matrix

And is estimated by the sample covariance matrix:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1q} \\ s_{21} & s_{11} & & \vdots \\ \vdots & & \ddots & \vdots \\ s_{q1} & \cdots & \cdots & s_{qq} \end{bmatrix} = \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T}{n-1}$$

Where $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{im} \end{bmatrix}$ is the i^{th} observation.



Correlation Matrix

The Covariance Matrix is commonly rescaled to be the correlation matrix:

$$P = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1q} \\ \rho_{21} & \rho_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{q1} & \cdots & \cdots & \rho_{qq} \end{bmatrix}$$

Where $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$



Regression in Terms of Matrices

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Sample Correlation Matrix

The Sample Correlation Matrix can be written as:

$$R = D^{-1/2}SD^{-1/2}$$

Where $D^{-1/2}$ is the matrix with $1/s_i$ on the diagonals.



What if Data is Missing?

- Missing Completely at Random?
- Imputation
- Multiple Imputation



(Multivariate) Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

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