Statistics 135

Chapter 3

Sample Geometry and Random Sampling

Chris Drake

Department of Statistics

University of California, Davis

Sample Mean, Covariance and Expectations

Recall data matrix, sample mean vector $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{S_n}$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{pmatrix}$$

note: column 1 contains the data on variable 1, column 2 contains the data on variable 2 etc, row one contains the measurements for subject 1, row 2 contains the measurements for subject 2 etc.

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \qquad \mathbf{S_n} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

1 assumption: the sample of observations $\mathbf{x_i}$ for $i = 1, .., n$ consists of n independent units with a common multivariate distribution $f(\mathbf{x_i}) = f(x_{1i}, .., x_{pi})$; ie we assume that subjects (sampling units) are sampled independently but the measurements taken on each unit are typically correlated.

2 fact: $E[\bar{\mathbf{X}}] = \mu$ therefore, $\bar{\mathbf{X}}$ is unbiased for μ

3 fact: the covariance of $\bar{X}$ is given by

$$Cov(\bar{\mathbf{X}}) = \frac{1}{n^2}\Big(\sum_{j=1}^{n} E\big((\mathbf{X_j} - \mu)(\mathbf{X_j} - \mu)'\big)\Big) = \frac{1}{n}\mathbf{\Sigma}$$

4 an unbiased estimate of $\mathbf{\Sigma}$ is given by $\mathbf{S} = (n/(n-1))\mathbf{S_n}$, recall that in $\mathbf{S_n}$ the diagonal and off-diagonal terms were divided by n; that estimator is biased.

5 the generalized sample variance is defined as $\mid \mathbf{S} \mid$, the determinant of $\mathbf{S}$.

6 if the sample covariance matrix is not of full rank (p), then the generalized sample variance is zero.

Sample statistics in matrix notation

1 Let $\mathbf{X}$ be the data matrix, then

$$\bar{\mathbf{x}} = \frac{1}{n}\ \mathbf{X}'\mathbf{l} \qquad\qquad since \qquad\qquad \mathbf{x}'\mathbf{l} = \sum_{i=1}^{n} x_i$$

note, the transpose of the data matrix contains the sample observations for variable x_1 in the first row, furthermore

$$\mathbf{l}\ \bar{\mathbf{x}}' = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{l}\mathbf{l}'\mathbf{X} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{pmatrix}$$

and

$$\mathbf{X} - \frac{1}{n}\mathbf{l}\mathbf{l}'\mathbf{X} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \dots & x_{p1} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{p2} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{pmatrix}$$

from this we get

$$\mathbf{S} = \frac{1}{n-1}\big(\mathbf{X} - \frac{1}{n}\mathbf{l}\mathbf{l}'\mathbf{X}\big)'\big(\mathbf{X} - \frac{1}{n}\mathbf{l}\mathbf{l}'\mathbf{X}\big) = \frac{1}{n-1}\mathbf{X}'\big(\mathbf{l} - \frac{1}{n}\mathbf{l}\mathbf{l}'\big)\mathbf{X}$$

if we let

$$\mathbf{D^{1/2}} = \begin{pmatrix} \sqrt{s_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{s_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{s_{pp}} \end{pmatrix}$$

and $\mathbf{D^{-1/2}}$ the diagonal matrix with $\sqrt{x_{ii}}$ replaced by $1/\sqrt{x_{ii}}$ and $\mathbf{R}$ is the sample correlation matrix with diagonal elements 1 and off-diagonal elements $s_{ik}/(\sqrt{s_{ii}}\sqrt{s_{kk}}$ then we can write

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2} \qquad and \qquad \mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$$

Linear combinations of random vectors

1 $\mathbf{c}$ a random vector and $\mathbf{c}'\mathbf{X}$ a linear combination with values $\mathbf{c}'\mathbf{x_j} = c_1 xj1 + c_2 x_{j2} + .. + c_n x_{jn}$ then

$$sample\ \ mean\ \ of\ \ \mathbf{c}'\mathbf{X} = \mathbf{c}'\bar{\mathbf{x}}$$

and

$$sample\ \ variance\ \ of\ \ \mathbf{c}'\mathbf{X} = \mathbf{c}'\mathbf{S}\mathbf{c}$$

2 if $\mathbf{b}$ is another random vector and $\mathbf{b}'\mathbf{X}$ another linear combination, then sample mean and variance are given by $\mathbf{b}'\bar{\mathbf{x}}$ and $\mathbf{b}'\mathbf{S}\mathbf{b}$ the sample covariance between $\mathbf{b}'\mathbf{X}$ and $\mathbf{c}'\mathbf{X}$ is given by

$$sample\ \ covariance\ \ = \mathbf{b}'\mathbf{S}\mathbf{c}$$