LECTURE 4

§6. RANDOM SAMPLING

With the vector concepts, we can now delve deeper into the geometrical interpretations of the descriptive statistics $\overline{x}$, S_n, and R. Many of our explanations use the representation of the columns of X as m vectors in n dimensions. We introduce the assumption that the observations constitute a random sample. Simply stated, random sampling implies that

(1) measurements taken on different items (or trials) are unrelated to one another; and

(2) the joint distribution of all m variables remains the same for all items.

Ultimately, it is this structure of the random sample that justifies a particular choice of distance and dictates the geometry for the n-dimensional representation of the data. Furthermore, when data can be treated as a random sample, statistical inferences are based on a solid foundation. Returning to geometric interpretations, we introduce a single number, called generalized variance, to describe variability. This generalization of variance is an integral part of the comparison of multivariate means. We use matrix algebra to provide concise expressions for the matrix products and sums that allow us to calculate $\overline{x}$ and S_n directly from the data matrix X. The connection between $\overline{x}$, S_n, and the means and covariances for linear combinations of variables is also clearly delineated, using the notion of matrix products.

§6.1. THE GEOMETRY OF THE SAMPLE

A single multivariate observation is the collection of measurements on m different variables taken on the same item or trial. If n observations have been obtained, the entire data set can be placed in an $n \times m$ matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots\dots & \dots\dots & \dots & \dots\dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

Each row of X represents a multivariate observation. Since the entire set of measurements is often one particular realization of what might have been observed, we say that the data are a sample of size n from a m-variate "population". The sample then consists of n measurements, each of which has m components.

The data can be plotted in two different ways. For the m-dimensional scatter plot, the rows of X represent n points in m-dimensional space. We can write

$$\mathbf{X} = \begin{bmatrix} x_{11}\ x_{12}\ ...\ x_{1m} \\ x_{21}\ x_{22}\ ...\ x_{2m} \\ .... \\ x_{n1}\ x_{n2}\ ...\ x_{nm} \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ .... \\ x'_n \end{bmatrix} \tag{2.1}$$

The row vector x'_j, representing the jth observation, contains the coordinates of a point.

The scatter plot of n points in m-dimensional space provides information on the locations and variability of the points. If the points are regarded as solid spheres, the sample mean vector $\overline{x}$, given by (1.6), is the center of balance. Variability occurs in more than one direction, and it is quantified by the sample variance-covariance matrix S_n. A single numerical measure of variability is provided by the determinant of the sample variance-covariance matrix. When m is greater than 3, this scatter plot representation cannot actually be graphed. Yet the consideration of the data as n points in m dimensions provides insights that are not readily available from algebraic expressions. Moreover, the concepts illustrated for $m = 2$ or $m = 3$ remain valid for the other cases.

Example 2.1. (Computing the mean vector) Compute the mean vector $\overline{x}$ from the data matrix.

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

The first point, x_1, has coordinates $x'_1 = [4, 1]$. Similarly, the remaining two points are $x'_2 = [-1, 3]$ and $x'_3 = [3, 5]$. Finally,

$$\overline{x} = \begin{bmatrix} \frac{4-1+3}{3} \\ \frac{1+3+5}{3} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

The alternative geometrical representation is constructed by considering the data as m vectors in n-dimensional space. Here we take the elements of the columns of the data matrix to be the coordinates of the vectors. Let

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots\dots & \dots\dots & \dots & \dots\dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} = [y_1 \vdots y_2 \vdots \dots \vdots y_m]. \tag{2.2}$$

Then the coordinates of the first point $y_1' = [x_{11}, x_{21}, \dots, x_{n1}]$ are the n measurements on the first variable. In general, the ith point $y_i' = [x_{1i}, x_{2i}, \cdots, x_{ni}]$ is determined by the n-tuple of all measurements on the ith variable. In this geometrical representation, we depict $y_1, \dots, y_m$ as vectors rather than points, as in the m-dimensional scatter plot.

Example 2.2 (Data as m vectors in n dimensions). Plot the following data as $m = 2$ vectors in $n = 3$ space:

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

Here $y_1' = [4, -1, 3]$ and $y_2' = [1, 3, 5]$.

Many of the algebraic expressions we will encounter in multivariate analysis can be related to the geometrical notions of length, angle, and volume. This is important because geometrical representations ordinarily facilitate understanding and lead to further insights.

Unfortunately, we are limited to visualizing objects in three dimensions, and consequently, the n-dimensional representation of the data matrix X may not seem like a particularly useful device for $n > 3$. It turns out, however, that geometrical relationships and the associated statistical concepts depicted for any three vectors remain valid regardless of their dimension. This follows because three vectors, even if n dimensional, can span no more than a three-dimensional space, just as two vectors with any number of components must lie in a plane. By selecting an appropriate three-dimensional perspective - that is, a portion of the n-dimensional space containing the three vectors of interest - a view is obtained that preserves both lengths and angles. Thus, it is possible, with the right choice of axes, to illustrate certain algebraic statistical concepts in terms of only two or three vectors of any dimension n. Since the specific choice of axes is not relevant to the geometry, we shall always label the coordinate axes 1,2, and 3.

It is possible to give a geometrical interpretation of the process of finding a sample mean. We start by defining the $n \times 1$ vector $1_n' = [1, 1, \ldots, 1]$. (To simplify the notation, the subscript n will be dropped when the dimension of the vector 1_n is clear from the context.) The vector 1 forms equal angles with each of the n coordinate axes, so the vector $(1/\sqrt{n})1$ has unit length in the equal-angle direction. Consider the vector $y_i' = [x_{1i}, x_{2i}, \ldots, x_{ni}]$. The projection of y_i on the unit vector $(1/\sqrt{n})1$ is

$$y_i' \left(\frac{1}{\sqrt{n}} \mathit{1} \right) \frac{1}{\sqrt{n}} 1 = \frac{x_{1i} + x_{2i} + \cdots + x_{ni}}{n} \mathit{1} = \overline{x}_i \mathit{1}. \tag{2.3}$$

That is, the sample mean $\overline{x}_i = (x_{1i} + x_{2i} + ... + x_{ni})/n = y_i' \mathit{1}/n$ corresponds to the multiple of *1* required to give the projection of y_i onto the line determined by *1*.
Further, for each y_i, we have the decomposition

$$y_i - \overline{x}_i \cdot 1,$$

where $\overline{x}_i \cdot 1$ is perpendicular to $y_i - \overline{x}_i \cdot 1$. The deviation, or mean corrected, vector is

$$d_i = y_i - \overline{x}_i \cdot 1 = \begin{bmatrix} x_{1i} - \overline{x}_i \\ x_{2i} - \overline{x}_i \\ \ldots\ldots\ldots \\ x_{ni} - \overline{x}_i \end{bmatrix}. \tag{2.4}$$

The elements of d_i are the deviations of the measurements on the i-th variable from their sample mean.

Example 2.3 (Decomposing a vector into its mean and deviation components).
Let us carry out the decomposition of y_i into $\overline{x}_i 1$ and $d_i = y_i - \overline{x}_i 1$, $i = 1, 2$, for the data given in Example 2.2:

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

Here $\overline{x}_1 = (4 - 1 + 3)/3 = 2$ and $\overline{x}_2 = (1 + 3 + 5)/3 = 3$, so

$$\overline{x}_1 1 = 2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}, \quad \overline{x}_2 1 = 3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix},$$

Consequently,

$$d_1 = y_1 - \overline{x}_1 1 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix},$$

and

$$d_2 = y_2 - \overline{x}_2\, 1 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}.$$

We note that $\overline{x}_1\, 1$ and $d_1 = y_1 - \overline{x}_1\, 1$ are perpendicular, because

$$(\overline{x}_1\, 1)'(y_1 - \overline{x}_1\, 1) = [2\ \ 2\ \ 2] \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = 4 - 6 + 2 = 0.$$

A similar result holds for $\overline{x}_2\, 1$ and $d_2 = y_2 - \overline{x}_2\, 1$. The decomposition is

$$y_1 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$y_2 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

For the time being, we are interested in the deviation (or residual) vectors $d_i = y_i - \overline{x}_i\, 1$. We have translated the deviation vectors to the origin without changing their lengths or orientations.

Now consider the squared lengths of the deviation vectors. We obtain

$$L^2_{d_i} = d_i'\, d_i = \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)^2 \tag{2.5}$$

(Length of deviation vector)^2 = sum of squared deviations.

Therefore, the squared length is proportional to the variance of the measurements on the ith variable. Equivalently, the length is proportional to the standard deviation. Longer vectors represent more variability than shorter vectors.

For any two deviation vectors d_i and d_k,

$$d_i'\, d_k = \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k). \tag{2.6}$$

Let θ_{ik} denote the angle formed by the vectors d_i and d_k. We get

$$d_i'\, d_k = L_{d_i}\, L_{d_k}\, \cos\theta_{ik}$$

or, using (2.5) and (2.6), we obtain

$$\sum_{j=1}^{n}(x_{ji}-\overline{x}_i)(x_{jk}-\overline{x}_k)=\sqrt{\sum_{j=1}^{n}(x_{ji}-\overline{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk}-\overline{x}_k)^2}\cos\theta_{ik},$$

so that

$$r_{ik}=\frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}=\cos\theta_{ik}. \qquad (2.7)$$

The cosine of the angle is the sample correlation coefficient. Thus, if the two deviation vectors have nearly the same orientation, the sample correlation will be close to 1. If the two vectors are nearly perpendicular, the sample correlation will be approximately zero. If the two vectors are oriented in nearly opposite directions, the sample correlation will be close to -1.

Example 2.4 (Calculating S_n and R from deviation vectors). Given the deviation vectors in Example 2.3, let us compute the sample variance-covariance matrix S_n and sample correlation matrix R using the geometrical concepts just introduced.

From Example 2.3,

$$d_1=\begin{bmatrix}2\\-3\\1\end{bmatrix}\quad\text{and}\quad d_2=\begin{bmatrix}-2\\0\\2\end{bmatrix}.$$

Now,

$$d_1'\,d_1=[2\ \ -3\ \ 1]\begin{bmatrix}2\\-3\\1\end{bmatrix}=14=\sum_{j=1}^{3}(x_{j1}-\overline{x}_1))^2=3s_{11}$$

or $s_{11}=14/3$. Also,

$$d_2'\,d_2=[-2\ \ 0\ \ 2]\begin{bmatrix}-2\\0\\2\end{bmatrix}=8=\sum_{j=1}^{3}(x_{j2}-\overline{x}_2))^2=3s_{22},$$

or $s_{22}=8/3$. Finally,

$$d_1'\,d_2=[2\ \ -3\ \ 1]\begin{bmatrix}-2\\0\\2\end{bmatrix}=-2=\sum_{j=1}^{3}(x_{j1}-\overline{x}_1))(x_{j2}-\overline{x}_2))=3s_{12}$$

or $s_{12}=-2/3$. Consequently,

$$r_{12}=\frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}}=\frac{-2/3}{\sqrt{14/3}\sqrt{8/3}}=-0.189$$

and

$$\mathbf{S}_n = \begin{bmatrix} 14/3 & -2/3 \\ -2/3 & 8/3 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1 & -0.189 \\ -0.189 & 1 \end{bmatrix}$$

The concepts of length, angle, and projection have provided us with a geometrical interpretation of the sample. We summarize as follows:

GEOMETRICAL INTERPRETATION OF THE SAMPLE

1. The projection of a column $\mathbf{y}_i$ of the data matrix $\mathbf{X}$ onto the equal angular vector 1 is the vector $\overline{x}_i \cdot 1$ has length $\sqrt{n}\,|\overline{x}_i|$. Therefore, the ith sample mean, $\overline{x}_i$, is related to the length of the projection of $\mathbf{y}_i$ on 1.

2. The information comprising $\mathbf{S}_n$ is obtained from the deviation vectors $d_i = y_i - \overline{x}_i \cdot 1 = [x_{1i} - \overline{x}_i, x_{2i} - \overline{x}_i, ..., x_{ni} - \overline{x}_i]'$. The square of the length of d_i is ns_{ii}, and the (inner) product between d_i and d_k is ns_{ik}[1].

3. The sample correlation r_{ik} is the cosine of the angle between d_i and d_k.

[1] The square of the length and the inner product are $(n-1)s_{ii}$ and $(n-1)s_{ik}$, respectively, when the divisor $n-1$ is used in the definitions of the sample variance and covariance.