Applied Statistics - Notes

260236

March 2024

Contents

1 Sample Geometry and Random Sampling

1.1 The Geometry of the Sample

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

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

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

Look at the matrix, n measurements (rows), each of which has p components (columns). In mathematics, each n row contains p columns and vice versa.

The data frame can be plotted in two different ways:

1. p-dimensional scatter plot, where the rows represent n points in p-dimensional space;
2. Geometrical representation, p vectors in n-dimensional space.

Scatter plot

For the **p-dimensional scatter plot**, the rows of $\mathbf{X}$ represent n points in p-dimensional space:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \begin{array}{l} \leftarrow \text{1st (multivariate) observation} \\ \\ \\ \leftarrow n\text{th (multivariate) observation} \end{array} \tag{2}$$

The row vector $\mathbf{x}_j'$, representing the jth observation, contains the coordinates of a point. The **scatter plot** of n points in p-dimensional space **provides information** on the **locations and variability of the points**.

Note: when p (dimensional space) is greater than 3, the **scatter plot** representation cannot actually be graphed. Yet the consideration of the data as n points in p dimensions provides **insights that are not readily available from algebraic expressions**.

Geometrical representation

The alternative **geometrical representation** is constructed by considering the data as p **vectors in n-dimensional space**. Here we take the elements of the columns of the data frame to be the coordinates of the vectors:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\mathbf{y}_1 \mid \mathbf{y}_2 \mid \cdots \mid \mathbf{y}_p] \tag{3}$$

Then the **coordinates** of the first point $\mathbf{y}_1 = [x_{11}, x_{21}, \ldots, x_{n1}]$ **are the** n **measurements** on the first variable.

In general, the ith point $\mathbf{y}_i = [x_{11}, x_{21}, \ldots, x_{n1}]$ is determined by the n-tuple of all measurements on the ith variable.

Geometrical representations usually **facilitate understanding** and lead to further insights. The ability to **relate algebraic expressions to the geometric concepts** of length, angle and volume is therefore **very important**.