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**.

Geometrical interpretation of the process of finding a sample mean

Before starting the explanation, you need to understand a few things.

- The **length** of a vector $\mathbf{x}' = [x_1, x_2, \ldots, x_n]$ with n components is defined by:

$$L_x = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{4}$$

  Multiplication of a vector $\mathbf{x}$ by a scalar c changes the length:

$$\begin{aligned} L_{cx} &= \sqrt{c^2 \cdot x_1^2 + c^2 \cdot x_2^2 + \cdots + c^2 \cdot x_n^2} \\ &= |c| \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \\ &= |c| \, L_x \end{aligned}$$

  So, for example, in $n = 2$ dimensions, the vector:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

  The length of $\mathbf{x}$, written L_x, is defined to be:

$$L_x = \sqrt{x_1^2 + x_2^2}$$

- Another important concept is **angle**. Consider two vectors in a plane and the angle θ between them: The value θ can be represented as the

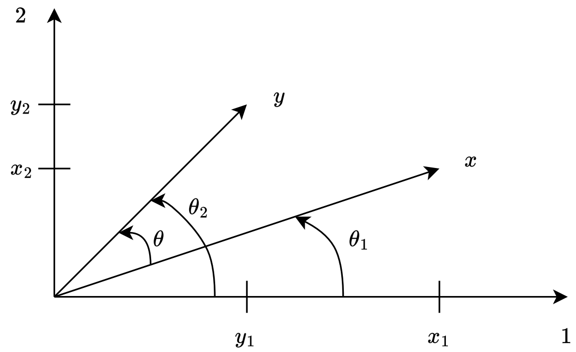


Figure 1: The angle θ between $\mathbf{x}' = [x_1, x_2]$ and $\mathbf{y}' = [y_1, y_2]$.

difference between the angles θ_1 and θ_2 formed by the two vectors and the first coordinate axis. Since, by definition:

$$\cos(\theta_1) = \frac{x_1}{L_x} \quad \cos(\theta_2) = \frac{y_1}{L_y}$$
$$\sin(\theta_1) = \frac{x_2}{L_x} \quad \sin(\theta_2) = \frac{y_2}{L_y}$$
$$\cos(\theta) = \cos(\theta_2 - \theta_1) = \cos(\theta_2)\cos(\theta_1) + \sin(\theta_2)\sin(\theta_1)$$

The angle θ between the two vectors $\mathbf{x}' = [x_1, x_2]$ and $\mathbf{y}' = [y_1, y_2]$ is specified by:

$$\cos(\theta) = \cos(\theta_2 - \theta_1) = \left(\frac{y_1}{L_y}\right)\left(\frac{x_1}{L_x}\right) + \left(\frac{y_2}{L_y}\right)\left(\frac{x^2}{L_x}\right) = \frac{x_1 y_1 + x_2 y_2}{L_x L_y} \tag{5}$$

- With the angle equation 5, it's convenient to introduce the **inner product** of two vectors:

$$\mathbf{x}\mathbf{y}' = x_1 y_1 + x_2 y_2$$

So let us rewrite:

  - The **length** equation 4:

$$\mathbf{x}'\mathbf{x} = x_1 x_1 + x_1 x_1 = x_1^2 + x_2^2 \longrightarrow L_x = \sqrt{x_1^2 + x_2^2} \Longrightarrow L_x = \sqrt{\mathbf{x}'\mathbf{x}} \tag{6}$$

  - The **angle** equation 5:

$$\cos(\theta) = \frac{x_1 y_1 + x_2 y_2}{L_x L_y} \Longrightarrow \cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{L_x L_y}$$

    And using the rewritten length equation:

$$\cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{L_x L_y} \Longrightarrow \cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}} \cdot \sqrt{\mathbf{y}'\mathbf{y}}}$$

- The **projection** (or shadown) of a vector $\mathbf{x}$ on a vector $\mathbf{y}$ is:

$$\frac{(\mathbf{x}'\mathbf{y})}{\mathbf{y}'\mathbf{y}}\mathbf{y} = \frac{(\mathbf{x}'\mathbf{y})}{L_y}\frac{1}{L_y}\mathbf{y} \tag{7}$$

Where the vector $\frac{1}{L_y}\mathbf{y}$ has unit length. The **length of the projection** is:

$$\frac{|\mathbf{x}'\mathbf{y}|}{L_y} = L_x\left|\frac{\mathbf{x}'\mathbf{y}}{L_x L_y}\right| = L_x\left|\cos\left(\theta\right)\right| \tag{8}$$

Where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$:

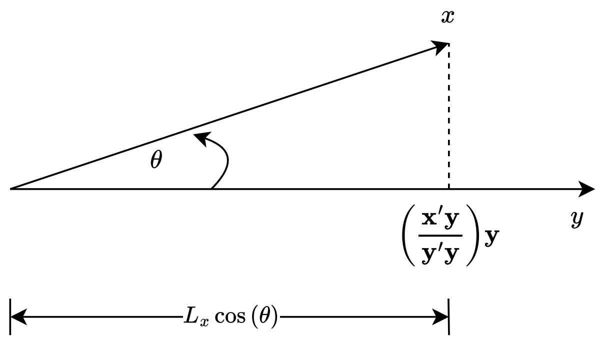


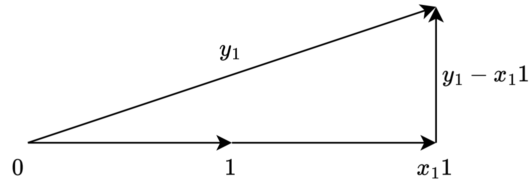
Figure 2: The projection of $\mathbf{x}$ on $\mathbf{y}$.

Start by defining the $n \times 1$ vector $\mathbf{1}_n' = [1, 1, \ldots, 1]$. The vector $\mathbf{1}$ forms equal angles with each of the n coordinates axes, so the vector $\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}$ has unit length in the equal-angle direction. Consider the vector $\mathbf{y}_i' = [x_{1i}, x_{2i}, \ldots, x_{ni}]$. The projection of $\mathbf{y}_i$ on the unit vector $\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}$ is:

$$\mathbf{y}_i'\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\frac{1}{\sqrt{n}}\mathbf{1} = \frac{x_{1i} + x_{2i} + \cdots + x_{ni}}{n}\mathbf{1} = \overline{x}_i\mathbf{1} \tag{9}$$

Although it may seem like a complex equation at first glance, it is nothing more than the mean! In fact, the **sample mean** $\overline{\mathbf{x}}_i = \frac{(x_{1i} + x_{2i} + \cdots + x_{ni})}{n} = \frac{\mathbf{y}_i'\mathbf{1}}{n}$ corresponds to the multiple of $\mathbf{1}$ required to give the projection of $\mathbf{y}_i$ onto the line determined by $\mathbf{1}$.

Furthermore, using the projection, you can obtain the **deviation** (**mean corrected**). For each $\mathbf{y}_i$ we have the decomposition:



Where $\overline{x}_i\mathbf{1}$ is perpendicular to $y_i - \overline{x}_i\mathbf{1}$. The **deviation**, or **mean corrected**, vector is:

$$\mathbf{d}_i = \mathbf{y}_i - \overline{x}_i\mathbf{1} = \begin{bmatrix} x_{1i} - \overline{x}_i \\ x_{2i} - \overline{x}_i \\ \vdots \\ x_{ni} - \overline{x}_i \end{bmatrix} \tag{10}$$

The **elements** of $\mathbf{d}_i$ are the **deviations of the measurements on the** i**th variable from their sample mean**.

Using the length rewritten with inner product (equation 6) and the deviation (equation 10), we obtain:

$$L^2_{\mathbf{d}_i} = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)^2 \tag{11}$$

$$(\text{Length of deviation vector})^2 = \text{sum of squared deviations}$$

From the sample standard deviation, we see that the **squared length is proportional to the variance** of the measurements on the ith variable. Equivalently, the **length is proportional to the standard deviation**. So longer vectors represent more variability than shorter vectors.

Furthermore, for any two deviation vectors $\mathbf{d}_i$ and $\mathbf{d}_k$:

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k) \tag{12}$$

And with a few mathematical operations, we can get it:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos(\theta_{ik}) \tag{13}$$

Where 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.