1.1 Introduction

Objectives of multivariate investigation

1. Data reduction or structural simplification
2. Sorting and Grouping
3. Investigation of the dependence among variables
4. Prediction
5. Hypothesis construction and Testing

1.3 The organization of data

Arrays of Basic Descriptive Statistics

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Sample Variances and covariances

$$S_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Covariances $s_{ij} = s_{ji}$

Variances

Symmetric Matrices

Sample Correlations

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

4 Nonlinear associations can exist that are not revealed by these descriptive statistics.

Possible Measurements of suspecting outliers

- Sum of squares of the deviations from the mean

$$W_{kk} = \sum_{j=1}^{n} (X_{jk} - \bar{X}_k)^2 \quad , K = 1, 2, \ldots, p$$

- Sum of cross product deviations

$$W_{ik} = \sum_{j=1}^{n} (X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k)$$

1.5 Distance

- most multivariate techniques are based on the simple concept of distance.

Distance from the origin to point $P = (X_1, X_2)$

$$d(O, P) = \sqrt{X_1^2 + X_2^2}$$

Distance from the origin to point $P = (X_1, X_2, \cdots X_p)$

$$d(O, P) = \sqrt{X_1^2 + X_2^2 + \cdots + X_p^2}$$

$$\Downarrow$$

$$d^2(O, P) = X_1^2 + X_2^2 + \cdots + X_p^2 = C^2$$

points equidistant from the origin

Equation of a hypersphere (circle if $p = 2$)

Euclidean distance is unsatisfactory for most statistical purpose because each coordinate contributes equally to the calculation of Euclidean distance. So instead, we use "statistical distance", which depend on correlations.

Statistical Distance :

- In case the observed variables share different scales or variation or both, it is important to use statistical distances instead of euclidean distances because the variable(s) with less variations or scale will have less effects on the analysis though they have the same effects. So, it is important to standardize every observation to center them to "0".

$$d(O,P) = \sqrt{\left(\frac{X_1}{\sqrt{S_{11}}}\right)^2 + \left(\frac{X_2}{\sqrt{S_{22}}}\right)^2} = \sqrt{\frac{X_1^2}{S_{11}} + \frac{X_2^2}{S_{22}}}$$

④ if the sample variances are the same and independent, euclidean distance is appropriate

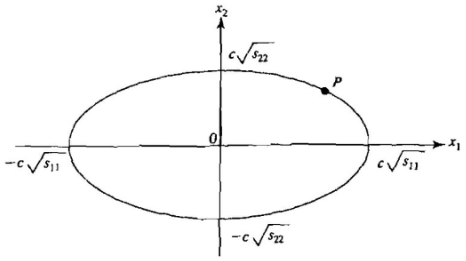


Figure 1.21 The ellipse of constant statistical distance $d^2(O, P) = x_1^2/s_{11} + x_2^2/s_{22} = c^2$.

The expression in (1-13) can be generalized to accommodate the calculation of statistical distance from an arbitrary point $P = (x_1, x_2)$ to any *fixed* point $Q = (y_1, y_2)$. If we assume that the coordinate variables vary independently of one another, the distance from P to Q is given by

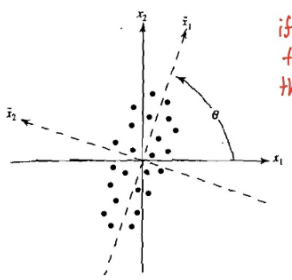
$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}} \tag{1-15}$$

The extension of this statistical distance to more than two dimensions is straightforward. Let the points P and Q have p coordinates such that $P = (x_1, x_2, \ldots, x_p)$ and $Q = (y_1, y_2, \ldots, y_p)$. Suppose Q is a fixed point [it may be the origin $O = (0, 0, \ldots, 0)$] and the coordinate variables vary independently of one another. Let $s_{11}, s_{22}, \ldots, s_{pp}$ be sample variances constructed from n measurements on $x_1, x_2, \ldots, x_p$, respectively. Then the statistical distance from P to Q is

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}} \tag{1-16}$$

- When variables' variations tend to depend on each other, the overall cluster of plots would be tilted. (When the variances are correlated).

At this point, we would have to rotate the original coordinate system through the angle θ while keeping the plots fixed.



if standardized scatter plot tend to be tilted, it is assumed that the variations are not independent.

Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.

forms

$$d(O,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{pp}x_p^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{p-1,p}x_{p-1}x_p} \tag{1-22}$$

and

$$d(P,Q) = \sqrt{\begin{aligned}&[a_{11}(x_1 - y_1)^2 + a_{22}(x_2 - y_2)^2 + \cdots + a_{pp}(x_p - y_p)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2)\\ &+ 2a_{13}(x_1 - y_1)(x_3 - y_3) + \cdots + 2a_{p-1,p}(x_{p-1} - y_{p-1})(x_p - y_p)]\end{aligned}} \tag{1-23}$$

a_{ij}'s are numbers such that the distances are always nonnegative.

rotating axes in case of $\dim(X) = 2$.

$$\tilde{X}_1 = X_1 \cos(\theta) + X_2 \sin(\theta)$$

$$\tilde{X}_2 = -X_1 \sin(\theta) + X_2 \cos(\theta)$$