L E C T U R E 1

§1. INTRODUCTION

Multivariate statistical analysis, or Multivariate Analysis, for short, is that branch of statistics which is devoted to the study of multivariate (or multidimensional) distributions and samples from those distributions. This, at least, is how the mathematical statistician would characterize this discipline.

We are interested in studying the interrelations among two or more variables, in looking for possible group differences in terms of these variables. Throughout the present course of lectures, we are going to be concerned with analyzing measurements made on several variables or characteristics. These measurements (commonly called **Data**) must frequently be arranged and displayed in various ways.

Multivariate data arise whenever an investigator, seeking to understand a phenomenon, selects a number $m \geq 1$ of variables or characters to record. We will use the notation x_{jk} to indicate the particular value of the k-th variable that is observed on the j-th trial. That is,

$$x_{jk} = \text{measurement of the } k\text{-th variable on the } j\text{-th trial.}$$

Consequently, n measurements on m variables can be displayed as a rectangular array (matrix), called $\mathbf{X}$, of n rows and m columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{lm} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2m} \\ \dots\dots & \dots\dots & \dots & \dots\dots & \dots & \dots\dots \\ x_{j1} & x_{j2} & \dots & x_{jk} & \dots & x_{jm} \\ \dots\dots & \dots\dots & \dots & \dots\dots & \dots & \dots\dots \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{nm} \end{bmatrix}$$

The matrix $\mathbf{X}$, then, contains the data consisting of all of the observations on all of the variables.

Considering data in the matrix form facilitates the exposition of the subject matter and allows numerical calculations to be performed in an orderly and efficient manner. The efficiency is twofold, as gains are attained in both

1) describing numerical calculations as operations on matrix

and

2) the implementation of the calculations on computers, which now use many languages and statistical packages to perform matrix operations.

§2. DESCRIPTIVE STATISTICS.

A large data set is bulky, and its very mass poses a serious to any attempt to visually extract pertinent information. Much of the information contained in the data can be assessed by calculating certain summary numbers, known as **descriptive statistics**. For example, the arithmetic average, or sample mean, is a descriptive statistic that provides a measure of location-that is, a "central value" for a set of numbers. And the average of the squares of the distances of all of the numbers from the mean provides a measure of the spread, or variation, in the numbers. We shall rely most heavily on descriptive statistics that measure location, variation, and linear association. The formal definitions of these quantities follow.

Let x_{11}, x_{21}, . . ., x_{n1} be n measurements on the first variable. Then the arithmetic average of these measurements is

$$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{j1}.$$

If the n measurements represent a subset of the full set of measurements that might have been observed, then $\bar{x}_1$ is also called the **sample mean** for the first variable. We adopt this terminology because our lectures are devoted to procedures designed for analyzing samples of measurements from larger collections. The sample mean can be computed from the n measurements on each of the m variables, so that, in general, there will be m sample means:

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, \qquad k = 1, 2, ..., m \tag{1.1}$$

A measure of spread is provided by the **sample variance**, defined for n measurements on the first variable as

$$s_1^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2,$$

where $\bar{x}_1$ is the sample mean of the x_{j1}'s, $j = 1, 2, ..., n$. In general, for m variables, we have

$$s_k^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2, \qquad k = 1, 2, ..., m. \tag{1.2}$$

Two comments are in order. First, many authors define the sample variance with a divisor of $n - 1$ rather than n. Later we will see that there are theoretical reasons for doing this, and it is particularly appropriate if the number of measurements, n, is small. Th two versions of the sample variance will always be differentiated by displaying the appropriate expression.

Second, although the s^2 notation is traditionally used to indicate the sample variance, we will eventually consider an array of quantities in which the sample variances lie along the main diagonal. In this situation, it is convenient to use double subscripts on the variances in order to indicate their positions in the matrix. Therefore, we introduce the notation s_{kk} to denote the same variance computed from measurements on the kth variable, and we have the notational identities

$$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2, \qquad k = 1, 2, ..., m. \tag{1.3}$$

The square root of the sample variance, $\sqrt{s_{kk}}$ is known as the **sample standard deviation**. This measure of variation is in the same units as the observations.

Consider n pairs of measurements on each of variables 1 and 2:

$$\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \qquad \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \quad ... \quad \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}.$$

That is, x_{j1} and x_{j2} are observed on the jth experiment, $j = 1, 2, ..., n$. A measure of linear association between the measurements of variables 1 and 2 is provided by the sample covariance

$$s_{12} = \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)$$

or the average product of the deviations from their respective means. If large values for one variable are observed in conjunction with large values for the other variable, and the small values also occur together, s_{12} will be positive. If large values from one

variable occur with small values for the other variable, s_{12} will be negative. If there is no particular association between the values for the two variables, s_{12} will be approximately zero.

The **sample covariance**

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k), \qquad i = 1, 2, ..., m, \quad k = 1, 2, ..., m, \tag{1.4}$$

measures the association between the ith and kth variables. We note that the covariance reduces to the sample variance when $i = k$. Moreover, $s_{ik} = s_{ki}$ for all i and k.

The final descriptive statistic considered here is the sample correlation coefficient. This measure of the linear association between two variables does not depend on the units of measurement. The sample correlation coefficient for the ith and kth variables is defined as

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}} \tag{1.5}$$

for $i = 1, 2, ..., m$ and $k = 1, 2, ..., m$. Note $r_{ik} = r_{ki}$ for all i and k.

The sample correlation coefficient is a standardized version of the sample covariance, where the product of the square roots of the sample variances provides the standardization. Note that r_{ik} has the same value whether n or $n - 1$ is chosen as the common divisor for s_{ii}, s_{kk} and s_{ik}.

The sample correlation coefficient r_{ik} can also be viewed as a sample covariance. Suppose the original values x_{ji} and x_{jk} are replaced by standardized values $\frac{x_{ji} - \bar{x}_i}{\sqrt{s_{ii}}}$ and $\frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}}$. The standardized values are commensurable because both sets are centered at zero and expressed in standard deviation units. The sample correlation coefficient is just the sample covariance of the standardized observations.

Although the signs of the sample correlation and the sample covariance are the same, the correlation is ordinarily easier to interpret because its magnitude is bounded. To summarize, the sample correlation r has the following properties:

1. The value of r must be between -1 and +1.

2. Here r measures the strength of the linear association. If $r = 0$, this implies a lack of linear association between the components. Otherwise, the sign of r indicates the direction of the association: $r < 0$ implies a tendency for one value in the pair to be larger than its average when the other is smaller than its average; and $r > 0$ implies a tendency for one value of the pair to be large when the other value is large and also for both values to be small together.

3. The value of r_{ik} remains unchanged if the measurements of the ith variable are changed to $y_{ji} = ax_{ji} + b$, $j = 1, 2, ..., n$, and the values of the kth variable are changed to $y_{jk} = cx_{jk} + d$, $j = 1, 2, ..., n$, provided that the constants a and c have the same sign.

The quantities s_{ik} and r_{ik} do not, in general, convey all there is to know about the association between two variables. Nonlinear associations can exist that are not revealed by these descriptive statistics. Covariance and correlation provide measures of linear association, or association along a line. Their values are less informative for other kinds of association. On the other hand, these quantities can be very sensitive to "wild" observations ("outliers") and may indicate association when, in fact, little exists. In spite of these shortcomings, covariance and correlation coefficients are routinely calculated and analyzed. They provide cogent numerical summaries of association when the data do not exhibit obvious nonlinear patterns of association and when wild observations are not present.

MATRICES OF BASIC DESCRIPTIVE STATISTICS

Sample means:

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ . \\ . \\ . \\ \bar{x}_m \end{bmatrix}$$

Sample variances and covariances:

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & ... & s_{1k} & ... & s_{1m} \\ s_{21} & s_{22} & ... & s_{2k} & ... & s_{2m} \\ ..... & ..... & ... & ..... & ... & .... \\ s_{j1} & s_{j2} & ... & s_{jk} & ... & s_{jm} \\ ..... & ..... & ... & ..... & ... & .... \\ s_{m1} & s_{m2} & ... & s_{mk} & ... & s_{mm} \end{bmatrix} \tag{1.6}$$

Sample correlations:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1k} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2k} & \ldots & r_{2m} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots & \ldots & \ldots\ldots \\ r_{m1} & r_{m2} & \ldots & r_{mk} & \ldots & 1 \end{bmatrix}$$

The sample mean matrix is denoted by $\bar{\mathbf{x}}$, the sample variance and covariance matrix by the capital letter $\mathbf{S}_n$, and the sample correlation matrix by $\mathbf{R}$. The subscript n on the matrix $\mathbf{S}_n$ is a mnemonic device used to remind you that n is employed as a divisor for the elements s_{jk}. The size of all of the matrices is determined by the number of variables, m.

The matrices $\mathbf{S}_n$ and $\mathbf{R}$ consist of m rows and m columns. The matrix $\bar{\mathbf{x}}$ is a single column with m rows. The first subscript on an entry in matrices $\mathbf{S}_n$ and $\mathbf{R}$ indicates the row; the second subscript indicates the column. Since $s_{ik} = s_{ki}$ and $r_{ik} = r_{ki}$ for all i and k, the entries in symmetric positions about the main northwest-southeast diagonals in matrices $\mathbf{S}_n$ and $\mathbf{R}$ are the same, and the matrices are said to be symmetric.

§3. COMBINING INDEPENDENT UNBIASED ESTIMATORS.

Let d_1 and d_2 denote independent unbiased estimators of θ, having known variances σ_1^2 and σ_2^2. That is, for $i = 1, 2$,

$$E[d_i] = \theta, \quad Var(d_i) = \sigma_i^2.$$

Any estimator of the form

$$d = \lambda d_1 + (1 - \lambda) d_2$$

will also be unbiased. To determine the value of λ that results in d having the smallest possible mean square error, note that

$$r(d, \theta) = Var(d) = \lambda^2 Var(d_1) + (1 - \lambda)^2 Var(d_2) = \lambda^2 \sigma_1^2 + (1 - \lambda)^2 \sigma_2^2.$$

Differentiation yields that

$$\frac{d}{d\lambda} r(d, \theta) = 2\lambda\sigma_1^2 - 2(1 - \lambda)\sigma_2^2.$$

To determine the value of λ that minimizes $r(d,\theta)$ – call it $\hat{\lambda}$ – set $\frac{d}{d\lambda}r(d,\theta)$ equal to 0 and solve for λ to obtain

$$2\hat{\lambda}\sigma_1^2 = 2(1-\hat{\lambda})\sigma_2^2$$

or

$$\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2+1/\sigma_2^2}.$$

In words, the optimal weight to give an estimator is inversely proportional to its variance (when all the estimators are unbiased and independent).

For an application of the foregoing, suppose that a conservation organization wants to determine the acidity content of a certain lake. To determine this quantity, they draw some water from the lake and then send samples of this water to n different laboratories. These laboratories will then, independently, test for acidity content by using their respective titration equipment, which is of differing precision. Specifically, suppose that d_i, the result of a titration test at laboratory i, is a random variable having mean θ, the true acidity of the sample water, and variance σ_i^2, $i = 1, ..., n$. If the quantities σ_i^2, $i = 1, .., n$ are known to the conservation organization, then they should estimate the acidity of the sampled water from the lake by

$$d = \frac{\sum_{i=1}^{n} d_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}.$$

The mean square error of d is as follows:

$$r(d,\theta) = Var(d) = \left(\sum_{i=1}^{n} 1/\sigma_i^2\right)^{-2} \sum_{i=1}^{n} \left(\frac{1}{\sigma_i^2}\right)^2 \sigma_i^2 = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2}.$$

A generalization of the result that the mean square error of an unbiased estimator is equal to its variance is that the mean square error of any estimator is equal to its variance plus the square of its bias. This follows since

$$r(d,\theta) = E[(d(\mathbf{X})-\theta)^2] = E[(d-E[d]+E[d]-\theta)^2] = E[(d-E[d])^2+(E[d]-\theta)^2+2(E[d]-\theta)(d-E[d])] =$$

$$= E[(d-E[d])^2] + E[(E[d]-\theta)^2] + 2E[(E[d]-\theta)(d-E[d])] = E[(d-E[d])^2] + (E[d]-\theta)^2.$$

The last equality follows since

$$E[d - E[d]] = 0.$$

Hence

$$r(d, \theta) = Var(d) + b_\theta^2(d).$$