# Stat 145 (Multivariate Statistics)

### Lesson 1.1 Introduction to Multivariate Analysis

## Learning Outcomes

1. Explain the nature of multivariate data.
2. Enumerate real-life applications of multivariate data analysis.
3. Compute the mean vector and variance-covariance matrix of multidimensional random variables.

### The Nature of Multivariate Data

Most real-world phenomenon are multi-faceted, meaning meaning they're influenced by a constellation of variables rather than a single factor. For example, educational outcomes, are not shaped solely by intelligence. They are influenced by socioeconomic status, school resources, teacher quality, student motivation, student health and nutrition status, and many more.

Often times, these variables are interacting with each other, and even some confound or mediate the relationship of others. These interrelationhships of many variables need to taken into consideration in the analysis. Hence,knowledge of statistical methods to handle multivariable data sets is inevitable.

The objectives of multivariate analysis fall into one of the following:

1. **Data reduction** or **structural simplification**. The phenomenon being studied is represented as simply as possible without sacrificing valuable information. It is hoped that this will make interpretation easier.
2. **Sorting** and **grouping**. Groups of "similar" objects or variables are created, based upon measured characteristics. Alternatively, rules for classifying objects into well-defined groups may be required.

3. **Investigation of the dependence among variables**. The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?
4. **Prediction**. Relationships between variables must be determined for the purpose of predicting the values of one or more variables on the basis of observations on the other variables.
5. **Hypothesis construction and testing**. Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

## Applications of Multivariate Techniques

### Data reduction or simplification

- Using data on several variables related to cancer patient responses to radiotherapy, a simple measure of patient response to radiotherapy was constructed.
- Track records from many nations were used to develop an index of performance for both male and female athletes.
- Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions.
- Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bea plants.
- A matrix of tactic similarities was developed from aggregate data derived from professional mediators. From this matrix the number of dimensions by which professional mediators judge the tactics they use in resolving disputes was determined.

### Sorting and grouping

- Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing (or planned) computer utilization.
- Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from nonalcoholics.
- Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease.
- The U.S. Internal Revenue Service uses data collected from tax returns to sort taxpayers into two groups: those that will be audited and those that will not.

**Investigation of the dependence among variables**

- Data on several variables were used to identify factors that were responsible for client success in hiring external consultants.
- Measurements of variables related to innovation, on the one hand, and variables related to the business environment and business organization, on the other hand, were used to discover why some firms are product innovators and some firms are not.
- Measurements of pulp fiber characteristics and subsequent measurements of characteristics of the paper made from them are used to examine the relations between pulp fiber properties and the resulting paper properties.The goal is to determine those fibers that lead to higher quality paper.
- The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance.

**Prediction**

- The associations between test scores, and several high school performance variables, and several college performance variables were used to develop predictors of success in college.
- Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments.
- Measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers.
- cDNA microarray experiments (gene expression data) are increasingly used to study the molecular variations among cancer tumors. A reliable classification of tumors is essential for successful diagnosis and treatment of cancer.

**Hypotheses testing**

- Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends.
- Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores.
- Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories.
- Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innovation.

## Structure of Multivariate Data

Multivariate data arise whenever an investigator, seeking to understand a social or physical phenomenon, selects a number of variables or characters to record. The values of these variables are all recorded for each distinct item, individual, or experimental unit. We will use the notation $x_{jk}$ to indicate the particular value of the $k^{th}$ variable that is observed on the $j^{th}$ item, or trial.

Multivariate data is typically represented as a matrix (array) where $j$ represents the row and $k$ the columns. For example,

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

Example 1.1

A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding numbers on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales): 42, 52, 48, 58 Variable 2 (number of books): 4, 5, 4, 3

Using the notation just introduced, we have

$x_{11} = 42$, $x_{21} = 52$, $x_{31} = 48$, $x_{41} = 58$

$x_{12} = 4$, $x_{22} = 5$, $x_{32} = 4$, $x_{42} = 3$

and the data matrix is

$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

with four rows and two columns.

## Descriptive Statistics

A large data set is bulky, and its very mass poses a serious obstacle 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*.

The sample mean can be computed from the $n$ measurements on each of the $p$ variables, so that, in general, there will be $p$ sample means:

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk},\ k = 1, 2, \cdots, p$$

The sample variances of the $p$ variables can be computed as

$$s_k^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2,\ k = 1, 2, \cdots, p$$

Meanwhile, the sample covariance between the $i^{th}$ and $k^{th}$ variables is calculated as

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k),\ i = 1, 2, \cdots, p;\ k = 1, 2, \cdots, p$$

Note that we can write $s_k^2 = s_{kk}$.

Finally, the sample correlation coefficient between the $i^{th}$ and $k^{th}$ variables is given by

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

The descriptive statistics computed from $n$ measurements on $p$ variables can also be organized into arrays/matrices.

**Sample Means**

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

**Sample variances and covariances**

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

**Sample correlations**

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

Example 1.2

Using the data in *Example 1.1*, verify that

$$\bar{\mathbf{x}} = \begin{bmatrix} 50 \\ 4 \end{bmatrix},$$

$$\mathbf{S} = \begin{bmatrix} 45.33 & -2 \\ -2 & 0.67 \end{bmatrix},$$

and

$$\mathbf{R} = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

Example 1.3:

Paper is manufactured in continuous sheets several feet wide. Because of the orientation of fibers within the paper, it has a different strength when measured in the direction produced by the machine than when measured across, or at right angles to, the machine direction. Below is a sample of 10 measurements of the following variables.

$x_1$ = density (grams per cubic centimeter)

$x_2$ = strength (pounds) in the machine direction

$x_3$ = strength 1pounds2 in the cross direction

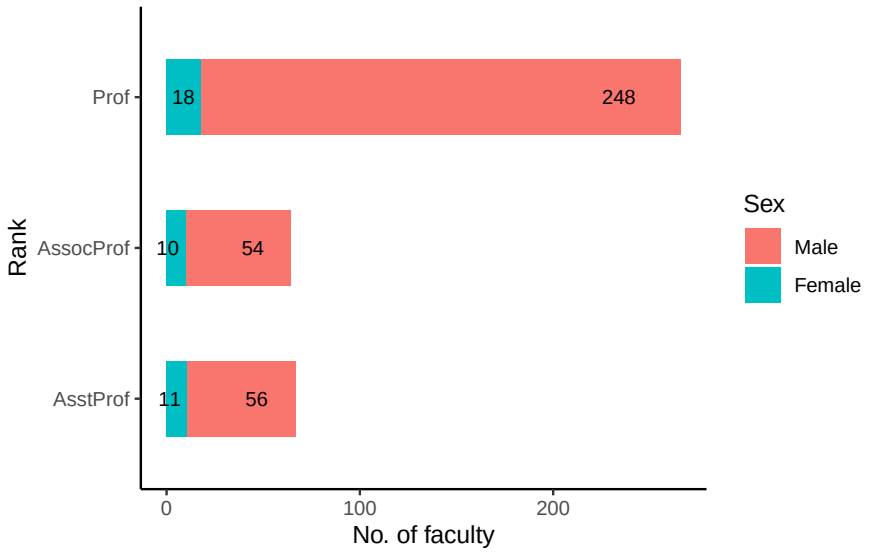
| Density | Machine direction | Cross direction |
|---|---|---|
| .801 | 121.41 | 70.42 |
| .824 | 127.70 | 72.47 |
| .841 | 129.20 | 78.20 |
| .816 | 131.80 | 74.89 |
| .840 | 135.10 | 71.21 |
| .842 | 131.50 | 78.39 |
| .820 | 126.70 | 69.02 |
| .802 | 115.10 | 73.10 |
| .828 | 130.80 | 79.28 |
| .819 | 124.60 | 76.48 |

Calculate the sample mean vector ($\bar{\mathbf{x}}$), the sample variance-covariance matrix ($\mathbf{S}$), and the sample correlation matrix ($\mathbf{R}$).

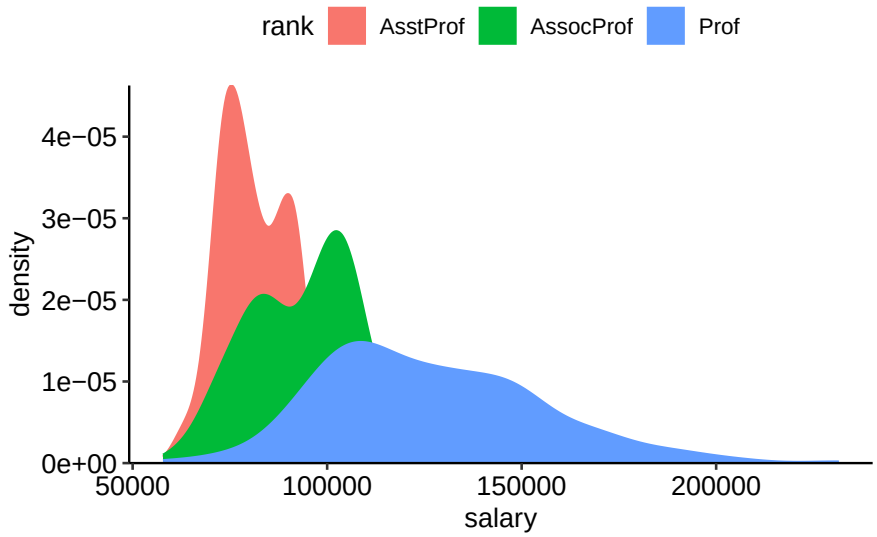
## Visualizing multivariate data

Just like in univariate data analysis, graphics play an important role to gain insights on the distribution and interrelationships between and among multiple variables. Some examples of multivariate data visualizations are shown here.

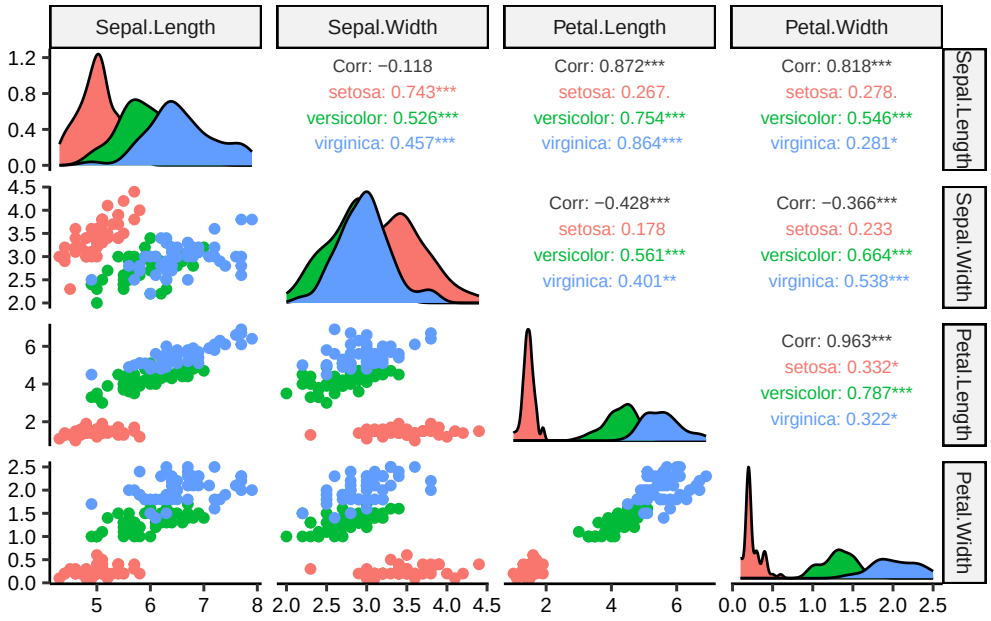
### Bar Plot

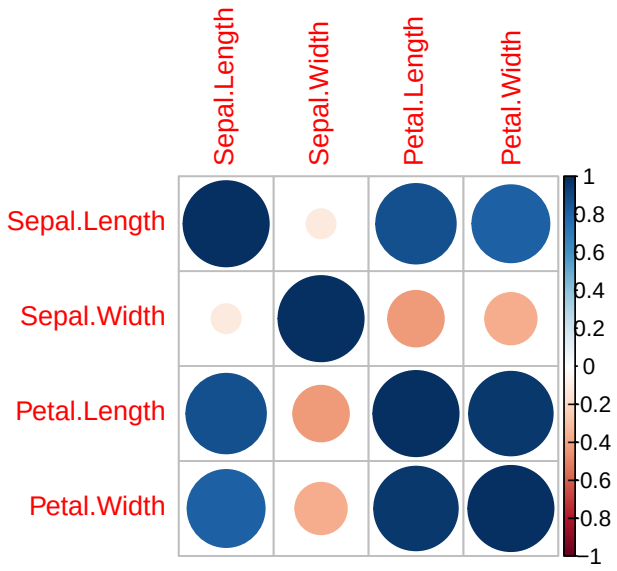


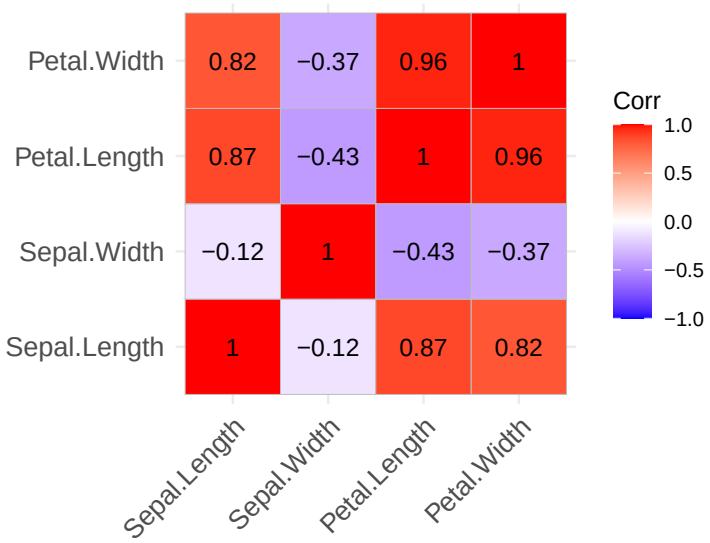
## Density Plot



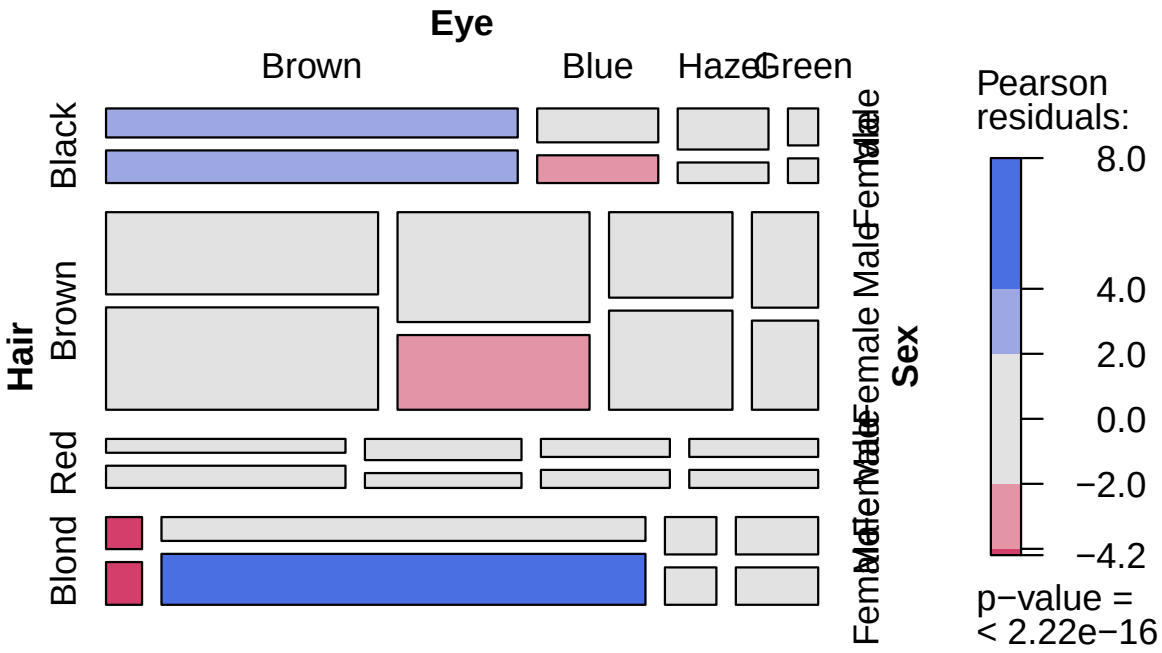

## Plots of Correlation Matrix




Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


Petal.Width
Petal.Length
Sepal.Width
Sepal.Length
0.82
−0.37
0.96
1
0.87
−0.43
1
0.96
−0.12
1
−0.43
−0.37
1
−0.12
0.87
0.82
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Corr
1.0
0.5
0.0
−0.5
−1.0

**Mosaic plot**


Eye
Brown
Blue
Hazel
Green
Hair
Black
Brown
Red
Blond
Sex
Male
Female
Pearson residuals:
8.0
4.0
2.0
0.0
−2.0
−4.2
p−value = < 2.22e−16