Chapter 1

ASPECTS OF MUTIVARIATE ANALYSIS

1.1 Introduction

Definition (Wikipedia): Multivariate analysis (MVA) is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical variable at a time.

The objectives of scientific investigations to which multivariate methods most naturally lend themselves include the following.

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

Examples: In the real world, most data collection schemes or designed experiments that provide data are multivariate in nature. Some examples of such situations are given below.

During a survey of households, several measurements on *each* household are taken. These measurements, being taken on the same household, will be dependent. For example, the education level of the head of the household and the annual income of the family are related.

During a production process, a number of different measurements such as the tensile strength, brittleness, diameter, etc. are taken on the same unit. Collectively such data are viewed as multivariate data.

Price of a car depends on several factors, say, year, mileage, warranty, HP, model, among many. Here year, mileage, warranty are correlated.

Body fitness depends on age, height, weight, amount of exercise, food habits etc. Here, height and weight are related.

A new drug is to be compared with a control for its effectiveness. Two different groups of patients are assigned to each of the two treatments and they are observed weekly for next two months. The periodic measurements on the same patient will exhibit dependence and thus the basic problem is multivariate in nature.

1.2 **Applications of Multivariate Techniques**

Some applications (among many) are describing below:

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

Read pages 3 and 4 for applications for each of the above categories.

1.2 The Organization of Data

Arrays

Multivariate data arise whenever an investigator, seeking to understand a social or physical phenomenon, selects a number $p \geq 1$ 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 notation x_{jk} to indicate the particular value of the *kth* variable that is observed on the *jth* item or trial. That is

x_{jk} =measurement of the *kth* variable on the *jth* item

Now n measurements on p variables can be displayed as follows

	Variable *1*	Variable *2*	….	Variable *k*	…	Variable *p*
Item 1	x_{11}	x_{12}	….	x_{1k}	…	x_{ip}
Item 2	x_{21}	x_{22}	….	x_{2k}	…	x_{2p}
…	…	…	….	…	…	…
Item *j*	x_{j1}	x_{j2}	….	x_{jk}	…	x_{jp}
…	…	…	…	…	…	
Item *n*	x_{n1}	x_{n2}	…	x_{nk}	…	x_{np}

These data can be displayed as a rectangular array, Called X, of *n* rows and *p* columns. The array X contains all of the observations on all of the variables.

$$X = \begin{bmatrix} x_{11}, & x_{12}, & \ldots . x_{1k}, \ldots . x_{1p} \\ x_{21}, & x_{22}, & \ldots x_{2k}, \ldots x_{2p} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ x_{j1} & x_{j2} \ldots, x_{jk} \ldots\ldots\ldots x_{jp} & \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ x_{n1}, & x_{n2}, & \ldots . x_{nk}, \ldots x_{np} \end{bmatrix}$$

Example 1.1, Page 6: Number of books and dollar sales

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 total amount of each sale and the number of books sold. The data are given below:

Variable 1 (dollars sales)	42 52 48 58
Variable 2 (# of books)	4 5 4 3

Then the data array X is (with 4 rows and 2 columns)

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

Here $x_{11}=42$, $x_{21}=52$, ……, $x_{42}=3$

Descriptive Statistics

Descriptive statistics describe the data. For example, mean, variance, standard deviation, correlations, skewness and kurtosis are descriptive statistics. We will discuss mostly discuss descriptive statistics that measure location, variation and linear association. The formal definitions of these quantities are given below.

Let $x_{11}, x_{12}, \cdots, x_{1n}$ be n measurements on variable 1. Then the sample mean of these measurements is

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

The second sample mean:

$$\bar{x}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{j2}$$

The p sample means:

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

The *sample variance* (which measures the variability of the data, also called *dispersion OR spread*) of n measurements for variable 1 is

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

The sample variance of *n* measurements for *p* variables

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

Note that, $s_k^2 = s_{kk}$ and the square root of the sample variance, $s_k = \sqrt{s_k^2} = \sqrt{s_{kk}}$ is known as the sample standard deviation (SD).

Note: Mostly we will be used SD to measure the variability as it has the same unit of measurement like as mean or median.

Sample Covariance: Consider *n* pairs of measurements on each of variables x_1 & x_2

$$\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \cdots, \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix},$$

A measure of linear association between the measurements of variables 1 and 2 is provided by the *sample covariance*. The sample covariance between variables x_1 and x_2 is denoted by s_{12} and defined as

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

The sample covariance between *ith* and *kth* variables is denoted by s_{ik} and defined as

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

This is the average product of the deviations from their respective means**.**

Sample correlation coefficient (also known as Pearson's product correlation coefficient)

The sample correlation coefficient between *ith* and *kth* variables is denoted by r_{ik} and defined as

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

$$= \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}}; \mathrm{i} = 1,2,..,\mathrm{p}$$

The sample correlation coefficient *r* has the following properties**:**

1. The values of *r* lie between -1 and +1 inclusive.
2. *r* measures the strength of linear association. Thus, *r=0* implies lack of linear association between two variables.

3. $r=\pm 1$, a perfect linear association.
4. $r > 0$ implies a tendency for one value of the pair to be large when other value is large and also both values to be small together.
5. $r < 0$ implies a tendency for one value in the pair to be large than its average when other value is smaller than its average
6. The value of r_{ik} remains unchanged if the measurements of the *ith* variable are changed to $y_{ji} = ax_{ji} + b$ and the values of the *kth* variable changed to $y_{jk} = cx_{jk} + b$ provided that the constants a and c have same sign. That means, *r* is invariant in both location and scale of measurements.

Arrays of Basic Descriptive Statistics

The descriptive statistics computed from *n* measurements on *p* variables can be organized into arrays.

Sample means

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

Sample variances and covariances

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

Sample correlation

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

Example 1.2, page 10

The arrays $\bar{X}$, S_n and R for bivariate data in **Example 1.1** are given below

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

$$S_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$$

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

r_{12}=*-0.36*, weak negative linear relationship between two variables X_1 and X_2.

Graphical Techniques:

Scatter Plot: Using SPSS we obtain the following scatter plot between variables 1 & 2.

Variable 1 (x_1):	3	4	2	6	8	2	5
Variable 2 (x_2):	5	5.5	4	7	10	5	7.5

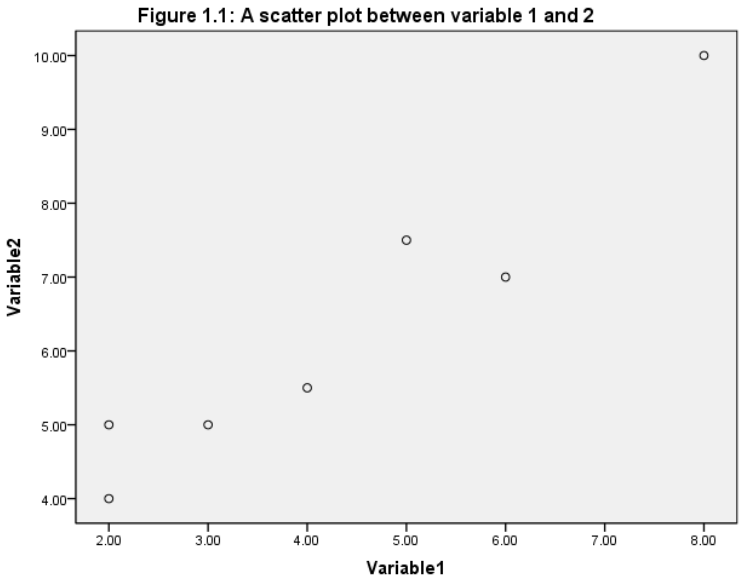


Figure 1.1: A scatter plot between variables x_1 and x_2

Using SPSS we obtain, $r_{12}=0.96$. Strong correlation between variables x_1 & x_2. The scatter diagram (Figure 1.1) gave the same impression about the strong linear relationship between variables x_1 & x_2.

Example 1.4 (A scatter plot for baseball data)

Table 1.1: 1977 Salary and Final Record for the National League East

Team	Player payroll (x_1)	Won-lost percentage (x_2)
Philadelphia	3497900	0.62
Pittsburg	2485475	0.59
St. Louis	1782875	0.51
Chicago	1725450	0.5
Montreal	1645575	0.46
New York	1469800	0.4

The scatter plot (using SPSS)

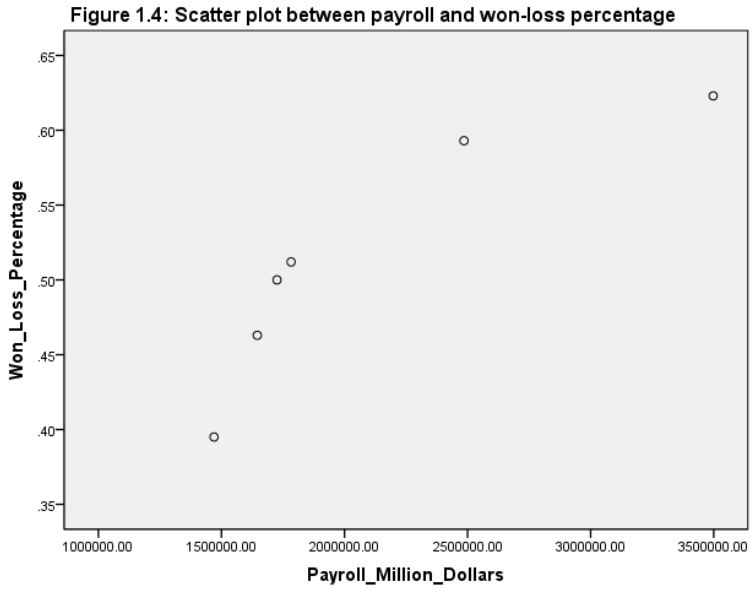


Figure 1.4: Salaries and won-lost percentage Table 1-1 (page 14)

Example 1.6, page 17: A zoologist obtained measurements on n=25 lizards. The weights or mass is given in grams while the snout-vent length (SVL) and hind limb span (HLS) are given in millimeters. The data are displayed in Table 1.3.

Table 1.3: Lizard Size Data

Lizard	Mass	SVL	HLS
1	5.5	59	113.5
2	10.4	75	142
3	9.2	69	124
4	9	67.5	125
5	7.1	62	129.5
6	6.6	62	123
7	11.3	74	140
8	2.4	47	97
9	15.5	86.5	162
10	9	69	126.5
11	8.2	70.5	136
12	6.6	64.5	116
13	7.6	67.5	135

14	10.1	73	136.5
15	10	73	135.5
16	10.1	77	139
17	7.6	61.5	118
18	7.73	66.5	133.5
19	12	79.5	150
20	10	74	137
21	5.1	59.5	116
22	9.2	68	123
23	12.1	75	141
24	7	66.5	117
24	6.9	63	117

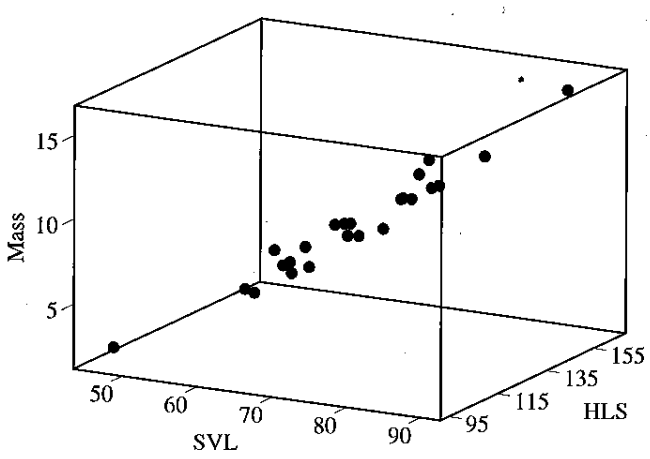
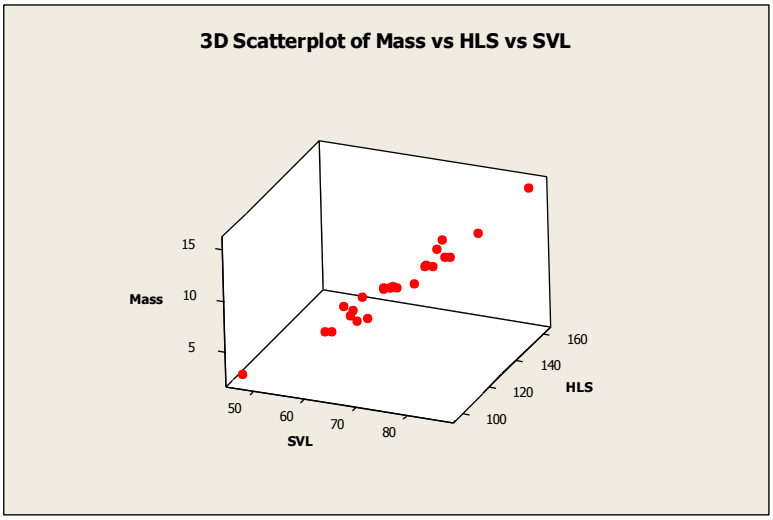


Figure 1.6 3D scatter plot of lizard data from Table 1.3.

Using Minitab



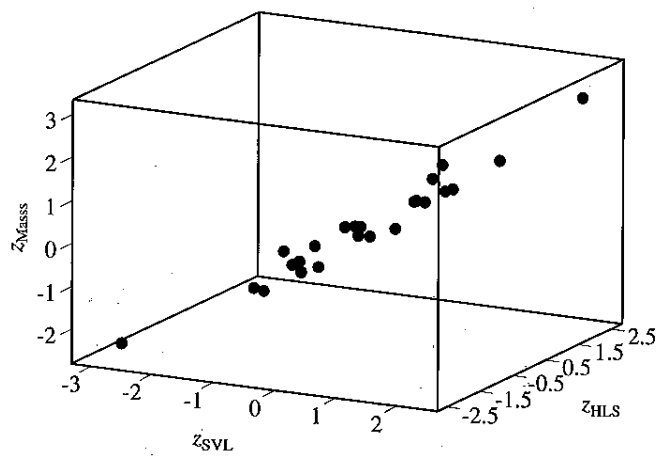


Figure 1.7 3D scatter plot of standardized lizard data. ■

From Figures 1.6 and 1.7, we can see that most of the variation is scatter about a one-dimensional straight line.

1.4 Data Display and Pictorial Representations

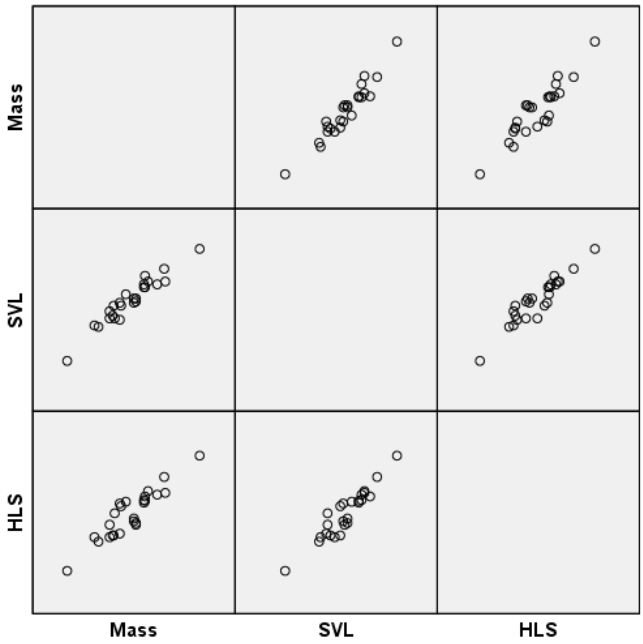
Consider data in **Example 1.6** and do a matrix plot, which is a linking multiple two-dimensional plots.

Correlation Matrix

Correlations

		Mass	SVL	HLS
Mass	Pearson Correlation	1	.961**	.916**
	Sig. (2-tailed)		.000	.000
	N	25	25	25
SVL	Pearson Correlation	.961**	1	.938**
	Sig. (2-tailed)	.000		.000
	N	25	25	25
HLS	Pearson Correlation	.916**	.938**	1
	Sig. (2-tailed)	.000	.000	
	N	25	25	25

*. Correlation is significant at the 0.01 level (2-tailed).



1.5 Distance

If we consider the point $P=(x_1, x_2)$ in the plane , the straight line distance, $d(O, P)$, from P to the origin $O=(0,0)$ is according to the Pythagorean theorem is given by

$$d(O,P) = \sqrt{x_1^2 + x_2^2}$$

The situation is illustrated in Figure 1.19 (page 30).

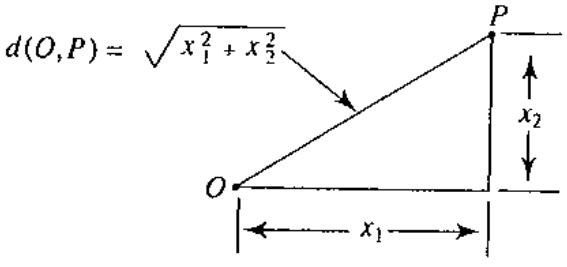


Figure 1.19 Distance given by the Pythagorean theorem.

In general, if the point P has p coordinates so that $P=(x_1, x_2, ...,x_p)$, the straight line distance from P to origin $O=(0,0,..,0)$ is

$$d(O,P)=\sqrt{x_1^2+x_2^2+\ldots+x_p^2}=\sqrt{x'x}=\|x\|$$

All points $(x_1, x_2 \ldots x_p)$ that lie a constant squared distance, such as c^2, from the origin satisfy the following equation

$$d(O,P)^2=x_1^2+x_2^2+\ldots+x_p^2=c^2$$

The straight line distance between two arbitrary points P and Q with coordinates $P=(x_1, x_2, ..,x_p)$, and $Q=(y_1, y_2,...,y_p)$, is given by

$$d(P,Q)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_p-y_p)^2}=\|x-y\|$$

Standardized Distance:

$$d(O,P)=\sqrt{(x_1^*)^2+(x_2^*)^2}$$
$$=\sqrt{\frac{x_1^2}{s_{11}}+\frac{x_2^2}{s_{22}}} \qquad (1\text{-}13)$$

Using (1-13), we see that all points which have coordinates (x_1, x_2) and are a constant squared distance c from the origin must satisfy

$$\frac{x_1^2}{s_{11}}+\frac{x_2^2}{s_{22}}=c^2 \qquad (1\text{-}14)$$

Equation (1-14) is the equation of an ellipse centered at the origin whose major and minor axes coincide with the coordinate axes. This general case is shown in Figure 1.21, page 32.

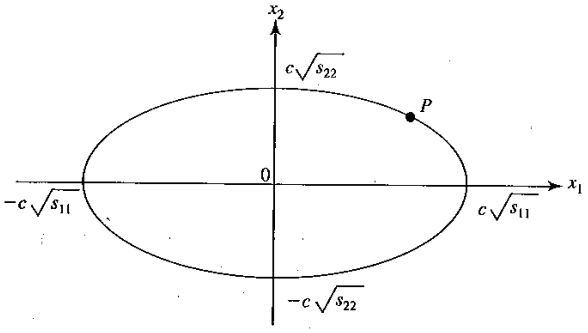


Figure 1.21 The ellipse of constant statistical distance $d^2(O, P) = x_1^2/s_{11} + x_2^2/s_{22} = c^2$.

Example 1.14, page 32: Calculating a statistical distance

$$d^2(O,P) = \frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = \frac{x_1^2}{4} + \frac{x_2^2}{1}$$

All points (x_1, x_2) that are a constant distance 1 from the origin satisfy the equation

$$\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$$

The coordinates of some points a unit distance from the origin are presented in the following Table

Coordinates: (x_1, x_2)	Distance: $\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$
(0, 1)	$\frac{0^2}{4} + \frac{1^1}{1} = 1$
(0, -1)	$\frac{0^2}{4} + \frac{(-1)^2}{1} = 1$
(2, 0)	$\frac{2^2}{4} + \frac{0}{1} = 1$
(1, $\sqrt{3}/2$)	$\frac{1^2}{4} + \frac{(\sqrt{3}/2)^2}{1} = 1$
(-2, 0)	$\frac{(-2)^2}{4} + \frac{(0)^2}{1} = 1$

A plot of the equation $\frac{x_1^2}{4}+\frac{x_2^2}{1}=1$ is given below

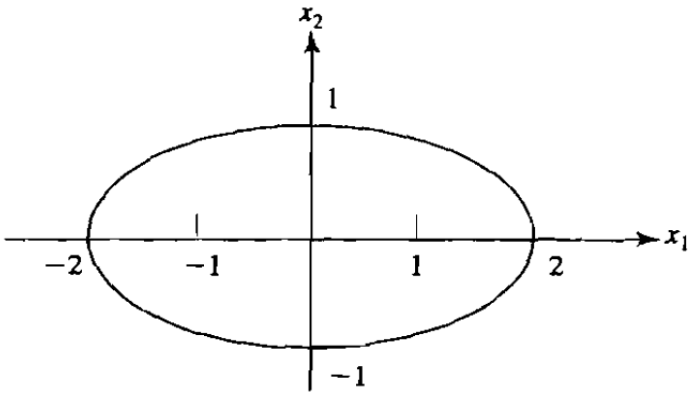


Figure 1.22 Ellipse of unit distance, $\frac{x_1^2}{4}+\frac{x_2^2}{1}=1$.

The expression in (1-13) can be generalized to accommodate the calculations 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 on one another, the distance from P to Q is given by

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

Let the points P and Q have p coordinated such that P=(x_1, x_2,...,x_p) and Q=(y_1, y_2, …,y_p). Suppose Q is a fixed point [it could be O=(0, 0, …, 0)] and the coordinate variables vary independently of one another. Let s_{11}, s_{22},….,s_{pp} be sample variances constructed from n measurements on x_1, x_2, …, x_p respectively. Then the statistical distance from P to Q is,

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