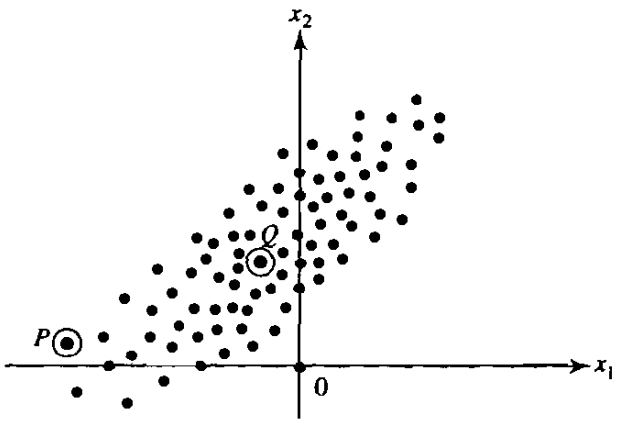


Figure 1.25 A cluster of points relative to a point P and the origin.

The need to consider statistical rather than Euclidean distance is illustrated heuristically in Figure 1.25. Figure 1.25 depicts a cluster of points whose center of gravity (sample mean) is indicated by the point Q. Consider the Euclidean distances from the point Q to the point P and the origin O. The Euclidean distance from Q to P is larger than the Euclidean distance from Q to O. However, P appears to be more like the points in the cluster than does the origin. If we take into account the variability of the points in the cluster and measure distance by the statistical distance in (1-20), then Q will be closer to P than to O. This result seems reasonable, given the nature of the scatter.

Other measures of distance can be advanced. (See Exercise 1.12.) At times, it is useful to consider distances that are not related to circles or ellipses. Any distance measure $d(P, Q)$ between two points P and Q is valid provided that it satisfies the following properties, where R is any other intermediate point:

$$\begin{aligned} &d(P, Q) = d(Q, P) \\ &d(P, Q) > 0 \text{ if } P \neq Q \\ &d(P, Q) = 0 \text{ if } P = Q \\ &d(P, Q) \leq d(P, R) + d(R, Q) \qquad \text{(triangle inequality)} \end{aligned} \tag{1-25}$$

1.6 Final Comments

We have attempted to motivate the study of multivariate analysis and to provide you with some rudimentary, but important, methods for organizing, summarizing, and displaying data. In addition, a general concept of distance has been introduced that will be used repeatedly in later chapters.

Exercises

1.1. Consider the seven pairs of measurements (x_1, x_2) plotted in Figure 1.1:

x_1	3	4	2	6	8	2	5
x_2	5	5.5	4	7	10	5	7.5

Calculate the sample means $\bar{x}_1$ and $\bar{x}_2$, the sample variances s_{11} and s_{22}, and the sample covariance s_{12}.

1.2. A morning newspaper lists the following used-car prices for a foreign compact with age x_1 measured in years and selling price x_2 measured in thousands of dollars:

x_1	1	2	3	3	4	5	6	8	9	11
x_2	18.95	19.00	17.95	15.54	14.00	12.95	8.94	7.49	6.00	3.99

(a) Construct a scatter plot of the data and marginal dot diagrams.

(b) Infer the sign of the sample covariance s_{12} from the scatter plot.

(c) Compute the sample means $\bar{x}_1$ and $\bar{x}_2$ and the sample variances s_{11} and s_{22}. Compute the sample covariance s_{12} and the sample correlation coefficient r_{12}. Interpret these quantities.

(d) Display the sample mean array $\bar{\mathbf{x}}$, the sample variance-covariance array $\mathbf{S}_n$, and the sample correlation array $\mathbf{R}$ using (1-8).

1.3. The following are five measurements on the variables x_1, x_2, and x_3:

x_1	9	2	6	5	8
x_2	12	8	6	4	10
x_3	3	4	0	2	1

Find the arrays $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$.

1.4. The world's 10 largest companies yield the following data:

The World's 10 Largest Companies[1]

Company	x_1 = sales (billions)	x_2 = profits (billions)	x_3 = assets (billions)
Citigroup	108.28	17.05	1,484.10
General Electric	152.36	16.59	750.33
American Intl Group	95.04	10.91	766.42
Bank of America	65.45	14.14	1,110.46
HSBC Group	62.97	9.52	1,031.29
ExxonMobil	263.99	25.33	195.26
Royal Dutch/Shell	265.19	18.54	193.83
BP	285.06	15.73	191.11
ING Group	92.01	8.10	1,175.16
Toyota Motor	165.68	11.13	211.15

[1]From www.Forbes.com partially based on *Forbes*The Forbes Global 2000, April 18, 2005.

(a) Plot the scatter diagram and marginal dot diagrams for variables x_1 and x_2. Comment on the appearance of the diagrams.

(b) Compute $\bar{x}_1$, $\bar{x}_2$, s_{11}, s_{22}, s_{12}, and r_{12}. Interpret r_{12}.

1.5. Use the data in Exercise 1.4.

(a) Plot the scatter diagrams and dot diagrams for (x_2, x_3) and (x_1, x_3). Comment on the patterns.

(b) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays for (x_1, x_2, x_3).

1.6. The data in Table 1.5 are 42 measurements on air-pollution variables recorded at 12:00 noon in the Los Angeles area on different days. (See also the air-pollution data on the web at www.prenhall.com/statistics.)

(a) Plot the marginal dot diagrams for all the variables.

(b) Construct the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays, and interpret the entries in $\mathbf{R}$.

Table 1.5 Air-Pollution Data

Wind (x_1)	Solar radiation (x_2)	CO (x_3)	NO (x_4)	NO_2 (x_5)	O_3 (x_6)	HC (x_7)
8	98	7	2	12	8	2
7	107	4	3	9	5	3
7	103	4	3	5	6	3
10	88	5	2	8	15	4
6	91	4	2	8	10	3
8	90	5	2	12	12	4
9	84	7	4	12	15	5
5	72	6	4	21	14	4
7	82	5	1	11	11	3
8	64	5	2	13	9	4
6	71	5	4	10	3	3
6	91	4	2	12	7	3
7	72	7	4	18	10	3
10	70	4	2	11	7	3
10	72	4	1	8	10	3
9	77	4	1	9	10	3
8	76	4	1	7	7	3
8	71	5	3	16	4	4
9	67	4	2	13	2	3
9	69	3	3	9	5	3
10	62	5	3	14	4	4
9	88	4	2	7	6	3
8	80	4	2	13	11	4
5	30	3	3	5	2	3
6	83	5	1	10	23	4
8	84	3	2	7	6	3
6	78	4	2	11	11	3
8	79	2	1	7	10	3
6	62	4	3	9	8	3
10	37	3	1	7	2	3
8	71	4	1	10	7	3
7	52	4	1	12	8	4
5	48	6	5	8	4	3
6	75	4	1	10	24	3
10	35	4	1	6	9	2
8	85	4	1	9	10	2
5	86	3	1	6	12	2
5	86	7	2	13	18	2
7	79	7	4	9	25	3
7	79	5	2	8	6	2
6	68	6	2	11	14	3
8	40	4	3	6	5	2

Source: Data courtesy of Professor G. C. Tiao.

1.7. You are given the following $n = 3$ observations on $p = 2$ variables:

$$\text{Variable 1:} \quad x_{11} = 2 \quad x_{21} = 3 \quad x_{31} = 4$$
$$\text{Variable 2:} \quad x_{12} = 1 \quad x_{22} = 2 \quad x_{32} = 4$$

(a) Plot the pairs of observations in the two-dimensional "variable space." That is, construct a two-dimensional scatter plot of the data.

(b) Plot the data as two points in the three-dimensional "item space."

1.8. Evaluate the distance of the point $P = (-1, -1)$ to the point $Q = (1, 0)$ using the Euclidean distance formula in (1-12) with $p = 2$ and using the statistical distance in (1-20) with $a_{11} = 1/3$, $a_{22} = 4/27$, and $a_{12} = 1/9$. Sketch the locus of points that are a constant squared statistical distance 1 from the point Q.

1.9. Consider the following eight pairs of measurements on two variables x_1 and x_2:

x_1	-6	-3	-2	1	2	5	6	8
x_2	-2	-3	1	-1	2	1	5	3

(a) Plot the data as a scatter diagram, and compute s_{11}, s_{22}, and s_{12}.

(b) Using (1-18), calculate the corresponding measurements on variables $\tilde{x}_1$ and $\tilde{x}_2$, assuming that the original coordinate axes are rotated through an angle of $\theta = 26°$ [given $\cos(26°) = .899$ and $\sin(26°) = .438$].

(c) Using the $\tilde{x}_1$ and $\tilde{x}_2$ measurements from (b), compute the sample variances $\tilde{s}_{11}$ and $\tilde{s}_{22}$.

(d) Consider the *new* pair of measurements $(x_1, x_2) = (4, -2)$. Transform these to measurements on $\tilde{x}_1$ and $\tilde{x}_2$ using (1-18), and calculate the distance $d(O, P)$ of the new point $P = (\tilde{x}_1, \tilde{x}_2)$ from the origin $O = (0, 0)$ using (1-17).
Note: You will need $\tilde{s}_{11}$ and $\tilde{s}_{22}$ from (c).

(e) Calculate the distance from $P = (4, -2)$ to the origin $O = (0, 0)$ using (1-19) and the expressions for a_{11}, a_{22}, and a_{12} in footnote 2.
Note: You will need s_{11}, s_{22}, and s_{12} from (a).
Compare the distance calculated here with the distance calculated using the $\tilde{x}_1$ and $\tilde{x}_2$ values in (d). (Within rounding error, the numbers should be the same.)

1.10. Are the following distance functions valid for distance from the origin? Explain.

(a) $x_1^2 + 4x_2^2 + x_1x_2 = (\text{distance})^2$

(b) $x_1^2 - 2x_2^2 = (\text{distance})^2$

1.11. Verify that distance defined by (1-20) with $a_{11} = 4$, $a_{22} = 1$, and $a_{12} = -1$ satisfies the first three conditions in (1-25). (The triangle inequality is more difficult to verify.)

1.12. Define the distance from the point $P = (x_1, x_2)$ to the origin $O = (0, 0)$ as

$$d(O, P) = \max(|x_1|, |x_2|)$$

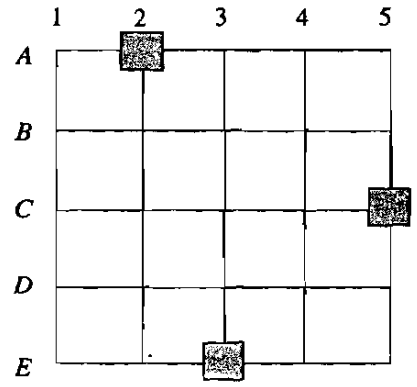
(a) Compute the distance from $P = (-3, 4)$ to the origin.

(b) Plot the locus of points whose squared distance from the origin is 1.

(c) Generalize the foregoing distance expression to points in p dimensions.

1.13. A large city has major roads laid out in a grid pattern, as indicated in the following diagram. Streets 1 through 5 run north–south (NS), and streets A through E run east–west (EW). Suppose there are retail stores located at intersections $(A, 2)$, $(E, 3)$, and $(C, 5)$.

Assume the distance along a street between two intersections in either the NS or EW direction is 1 unit. Define the distance between any two intersections (points) on the grid to be the "city block" distance. [For example, the distance between intersections $(D, 1)$ and $(C, 2)$, which we might call $d((D,1),(C,2))$, is given by $d((D,1),(C,2)) = d((D,1),(D,2)) + d((D,2),(C,2)) = 1 + 1 = 2$. Also, $d((D,1),(C,2)) = d((D,1),(C,1)) + d((C,1),(C,2)) = 1 + 1 = 2$.]



Locate a supply facility (warehouse) at an intersection such that the sum of the distances from the warehouse to the three retail stores is minimized.

The following exercises contain fairly extensive data sets. A computer may be necessary for the required calculations.

1.14. Table 1.6 contains some of the raw data discussed in Section 1.2. (See also the multiple-sclerosis data on the web at www.prenhall.com/statistics.) Two different visual stimuli ($S1$ and $S2$) produced responses in both the left eye (L) and the right eye (R) of subjects in the study groups. The values recorded in the table include x_1 (subject's age); x_2 (total response of both eyes to stimulus $S1$, that is, $S1L + S1R$); x_3 (difference between responses of eyes to stimulus $S1$, $|S1L - S1R|$); and so forth.

(a) Plot the two-dimensional scatter diagram for the variables x_2 and x_4 for the multiple-sclerosis group. Comment on the appearance of the diagram.

(b) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays for the non-multiple-sclerosis and multiple-sclerosis groups separately.

1.15. Some of the 98 measurements described in Section 1.2 are listed in Table 1.7 (See also the radiotherapy data on the web at www.prenhall.com/statistics.) The data consist of average ratings over the course of treatment for patients undergoing radiotherapy. Variables measured include x_1 (number of symptoms, such as sore throat or nausea); x_2 (amount of activity, on a 1–5 scale); x_3 (amount of sleep, on a 1–5 scale); x_4 (amount of food consumed, on a 1–3 scale); x_5 (appetite, on a 1–5 scale); and x_6 (skin reaction, on a 0–3 scale).

(a) Construct the two-dimensional scatter plot for variables x_2 and x_3 and the marginal dot diagrams (or histograms). Do there appear to be any errors in the x_3 data?

(b) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Interpret the pairwise correlations.

1.16. At the start of a study to determine whether exercise or dietary supplements would slow bone loss in older women, an investigator measured the mineral content of bones by photon absorptiometry. Measurements were recorded for three bones on the dominant and nondominant sides and are shown in Table 1.8. (See also the mineral-content data on the web at www.prenhall.com/statistics.)

Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Interpret the pairwise correlations.

Table 1.6 Multiple-Sclerosis Data

Non-Multiple-Sclerosis Group Data

Subject number	x_1 (Age)	x_2 $(S1L + S1R)$	x_3 $\lvert S1L - S1R \rvert$	x_4 $(S2L + S2R)$	x_5 $\lvert S2L - S2R \rvert$
1	18	152.0	1.6	198.4	.0
2	19	138.0	.4	180.8	1.6
3	20	144.0	.0	186.4	.8
4	20	143.6	3.2	194.8	.0
5	20	148.8	.0	217.6	.0
⋮	⋮	⋮	⋮	⋮	⋮
65	67	154.4	2.4	205.2	6.0
66	69	171.2	1.6	210.4	.8
67	73	157.2	.4	204.8	.0
68	74	175.2	5.6	235.6	.4
69	79	155.0	1.4	204.4	.0

Multiple-Sclerosis Group Data

Subject number	x_1	x_2	x_3	x_4	x_5
1	23	148.0	.8	205.4	.6
2	25	195.2	3.2	262.8	.4
3	25	158.0	8.0	209.8	12.2
4	28	134.4	.0	198.4	3.2
5	29	190.2	14.2	243.8	10.6
⋮	⋮	⋮	⋮	⋮	⋮
25	57	165.6	16.8	229.2	15.6
26	58	238.4	8.0	304.4	6.0
27	58	164.0	.8	216.8	.8
28	58	169.8	.0	219.2	1.6
29	59	199.8	4.6	250.2	1.0

Source: Data courtesy of Dr. G. G. Celesia.

Table 1.7 Radiotherapy Data

x_1 Symptoms	x_2 Activity	x_3 Sleep	x_4 Eat	x_5 Appetite	x_6 Skin reaction
.889	1.389	1.555	2.222	1.945	1.000
2.813	1.437	.999	2.312	2.312	2.000
1.454	1.091	2.364	2.455	2.909	3.000
.294	.941	1.059	2.000	1.000	1.000
2.727	2.545	2.819	2.727	4.091	.000
⋮	⋮	⋮	⋮	⋮	⋮
4.100	1.900	2.800	2.000	2.600	2.000
.125	1.062	1.437	1.875	1.563	.000
6.231	2.769	1.462	2.385	4.000	2.000
3.000	1.455	2.090	2.273	3.272	2.000
.889	1.000	1.000	2.000	1.000	2.000

Source: Data courtesy of Mrs. Annette Tealey, R.N. Values of x_2 and x_3 less than 1.0 are due to errors in the data-collection process. Rows containing values of x_2 and x_3 less than 1.0 may be omitted.

Table 1.8 Mineral Content in Bones

Subject number	Dominant radius	Radius	Dominant humerus	Humerus	Dominant ulna	Ulna
1	1.103	1.052	2.139	2.238	.873	.872
2	.842	.859	1.873	1.741	.590	.744
3	.925	.873	1.887	1.809	.767	.713
4	.857	.744	1.739	1.547	.706	.674
5	.795	.809	1.734	1.715	.549	.654
6	.787	.779	1.509	1.474	.782	.571
7	.933	.880	1.695	1.656	.737	.803
8	.799	.851	1.740	1.777	.618	.682
9	.945	.876	1.811	1.759	.853	.777
10	.921	.906	1.954	2.009	.823	.765
11	.792	.825	1.624	1.657	.686	.668
12	.815	.751	2.204	1.846	.678	.546
13	.755	.724	1.508	1.458	.662	.595
14	.880	.866	1.786	1.811	.810	.819
15	.900	.838	1.902	1.606	.723	.677
16	.764	.757	1.743	1.794	.586	.541
17	.733	.748	1.863	1.869	.672	.752
18	.932	.898	2.028	2.032	.836	.805
19	.856	.786	1.390	1.324	.578	.610
20	.890	.950	2.187	2.087	.758	.718
21	.688	.532	1.650	1.378	.533	.482
22	.940	.850	2.334	2.225	.757	.731
23	.493	.616	1.037	1.268	.546	.615
24	.835	.752	1.509	1.422	.618	.664
25	.915	.936	1.971	1.869	.869	.868

Source: Data courtesy of Everett Smith.

1.17. Some of the data described in Section 1.2 are listed in Table 1.9. (See also the national-track-records data on the web at www.prenhall.com/statistics.) The national track records for women in 54 countries can be examined for the relationships among the running events. Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Notice the magnitudes of the correlation coefficients as you go from the shorter (100-meter) to the longer (marathon) running distances. Interpret these pairwise correlations.

1.18. Convert the national track records for women in Table 1.9 to speeds measured in meters per second. For example, the record speed for the 100-m dash for Argentinian women is 100 m/11.57 sec = 8.643 m/sec. Notice that the records for the 800-m, 1500-m, 3000-m and marathon runs are measured in minutes. The marathon is 26.2 miles, or 42,195 meters, long. Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Notice the magnitudes of the correlation coefficients as you go from the shorter (100 m) to the longer (marathon) running distances. Interpret these pairwise correlations. Compare your results with the results you obtained in Exercise 1.17.

1.19. Create the scatter plot and boxplot displays of Figure 1.5 for (a) the mineral-content data in Table 1.8 and (b) the national-track-records data in Table 1.9.

Table 1.9 National Track Records for Women

Country	100 m (s)	200 m (s)	400 m (s)	800 m (min)	1500 m (min)	3000 m (min)	Marathon (min)
Argentina	11.57	22.94	52.50	2.05	4.25	9.19	150.32
Australia	11.12	22.23	48.63	1.98	4.02	8.63	143.51
Austria	11.15	22.70	50.62	1.94	4.05	8.78	154.35
Belgium	11.14	22.48	51.45	1.97	4.08	8.82	143.05
Bermuda	11.46	23.05	53.30	2.07	4.29	9.81	174.18
Brazil	11.17	22.60	50.62	1.97	4.17	9.04	147.41
Canada	10.98	22.62	49.91	1.97	4.00	8.54	148.36
Chile	11.65	23.84	53.68	2.00	4.22	9.26	152.23
China	10.79	22.01	49.81	1.93	3.84	8.10	139.39
Columbia	11.31	22.92	49.64	2.04	4.34	9.37	155.19
Cook Islands	12.52	25.91	61.65	2.28	4.82	11.10	212.33
Costa Rica	11.72	23.92	52.57	2.10	4.52	9.84	164.33
Czech Republic	11.09	21.97	47.99	1.89	4.03	8.87	145.19
Denmark	11.42	23.36	52.92	2.02	4.12	8.71	149.34
Dominican Republic	11.63	23.91	53.02	2.09	4.54	9.89	166.46
Finland	11.13	22.39	50.14	2.01	4.10	8.69	148.00
France	10.73	21.99	48.25	1.94	4.03	8.64	148.27
Germany	10.81	21.71	47.60	1.92	3.96	8.51	141.45
Great Britain	11.10	22.10	49.43	1.94	3.97	8.37	135.25
Greece	10.83	22.67	50.56	2.00	4.09	8.96	153.40
Guatemala	11.92	24.50	55.64	2.15	4.48	9.71	171.33
Hungary	11.41	23.06	51.50	1.99	4.02	8.55	148.50
India	11.56	23.86	55.08	2.10	4.36	9.50	154.29
Indonesia	11.38	22.82	51.05	2.00	4.10	9.11	158.10
Ireland	11.43	23.02	51.07	2.01	3.98	8.36	142.23
Israel	11.45	23.15	52.06	2.07	4.24	9.33	156.36
Italy	11.14	22.60	51.31	1.96	3.98	8.59	143.47
Japan	11.36	23.33	51.93	2.01	4.16	8.74	139.41
Kenya	11.62	23.37	51.56	1.97	3.96	8.39	138.47
Korea, South	11.49	23.80	53.67	2.09	4.24	9.01	146.12
Korea, North	11.80	25.10	56.23	1.97	4.25	8.96	145.31
Luxembourg	11.76	23.96	56.07	2.07	4.35	9.21	149.23
Malaysia	11.50	23.37	52.56	2.12	4.39	9.31	169.28
Mauritius	11.72	23.83	54.62	2.06	4.33	9.24	167.09
Mexico	11.09	23.13	48.89	2.02	4.19	8.89	144.06
Myanmar(Burma)	11.66	23.69	52.96	2.03	4.20	9.08	158.42
Netherlands	11.08	22.81	51.35	1.93	4.06	8.57	143.43
New Zealand	11.32	23.13	51.60	1.97	4.10	8.76	146.46
Norway	11.41	23.31	52.45	2.03	4.01	8.53	141.06
Papua New Guinea	11.96	24.68	55.18	2.24	4.62	10.21	221.14
Philippines	11.28	23.35	54.75	2.12	4.41	9.81	165.48
Poland	10.93	22.13	49.28	1.95	3.99	8.53	144.18
Portugal	11.30	22.88	51.92	1.98	3.96	8.50	143.29
Romania	11.30	22.35	49.88	1.92	3.90	8.36	142.50
Russia	10.77	21.87	49.11	1.91	3.87	8.38	141.31
Samoa	12.38	25.45	56.32	2.29	5.42	13.12	191.58

(continues)

Country	100 m (s)	200 m (s)	400 m (s)	800 m (min)	1500 m (min)	3000 m (min)	Marathon (min)
Singapore	12.13	24.54	55.08	2.12	4.52	9.94	154.41
Spain	11.06	22.38	49.67	1.96	4.01	8.48	146.51
Sweden	11.16	22.82	51.69	1.99	4.09	8.81	150.39
Switzerland	11.34	22.88	51.32	1.98	3.97	8.60	145.51
Taiwan	11.22	22.56	52.74	2.08	4.38	9.63	159.53
Thailand	11.33	23.30	52.60	2.06	4.38	10.07	162.39
Turkey	11.25	22.71	53.15	2.01	3.92	8.53	151.43
U.S.A.	10.49	21.34	48.83	1.94	3.95	8.43	141.16

Source: *IAAF/ATFS Track and Field Handbook for Helsinki* 2005 (courtesy of Ottavio Castellini).

1.20. Refer to the bankruptcy data in Table 11.4, page 657, and on the following website www.prenhall.com/statistics. Using appropriate computer software,

(a) View the entire data set in x_1, x_2, x_3 space. Rotate the coordinate axes in various directions. Check for unusual observations.

(b) Highlight the set of points corresponding to the bankrupt firms. Examine various three-dimensional perspectives. Are there some orientations of three-dimensional space for which the bankrupt firms can be distinguished from the nonbankrupt firms? Are there observations in each of the two groups that are likely to have a significant impact on any rule developed to classify firms based on the sample means, variances, and covariances calculated from these data? (See Exercise 11.24.)

1.21. Refer to the milk transportation-cost data in Table 6.10, page 345, and on the web at www.prenhall.com/statistics. Using appropriate computer software,

(a) View the entire data set in three dimensions. Rotate the coordinate axes in various directions. Check for unusual observations.

(b) Highlight the set of points corresponding to gasoline trucks. Do any of the gasoline-truck points appear to be multivariate outliers? (See Exercise 6.17.) Are there some orientations of x_1, x_2, x_3 space for which the set of points representing gasoline trucks can be readily distinguished from the set of points representing diesel trucks?

1.22. Refer to the oxygen-consumption data in Table 6.12, page 348, and on the web at www.prenhall.com/statistics. Using appropriate computer software,

(a) View the entire data set in three dimensions employing various combinations of three variables to represent the coordinate axes. Begin with the x_1, x_2, x_3 space.

(b) Check this data set for outliers.

1.23. Using the data in Table 11.9, page 666, and on the web at www.prenhall.com/statistics, represent the cereals in each of the following ways.

(a) Stars.

(b) Chernoff faces. (Experiment with the assignment of variables to facial characteristics.)

1.24. Using the utility data in Table 12.4, page 688, and on the web at www.prenhall.com/statistics, represent the public utility companies as Chernoff faces with assignments of variables to facial characteristics different from those considered in Example 1.12. Compare your faces with the faces in Figure 1.17. Are different groupings indicated?

1.25. Using the data in Table 12.4 and on the web at www.prenhall.com/statistics, represent the 22 public utility companies as stars. Visually group the companies into four or five clusters.

1.26. The data in Table 1.10 (see the bull data on the web at www.prenhall.com/statistics) are the measured characteristics of 76 young (less than two years old) bulls sold at auction. Also included in the table are the selling prices (SalePr) of these bulls. The column headings (variables) are defined as follows:

$$\text{Breed} = \begin{cases} 1 & \text{Angus} \\ 5 & \text{Hereford} \\ 8 & \text{Simental} \end{cases}$$

YrHgt = Yearling height at shoulder (inches)

FtFrBody = Fat free body (pounds)

PrctFFB = Percent fat-free body

Frame = Scale from 1 (small) to 8 (large)

BkFat = Back fat (inches)

SaleHt = Sale height at shoulder (inches)

SaleWt = Sale weight (pounds)

(a) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Interpret the pairwise correlations. Do some of these variables appear to distinguish one breed from another?

(b) View the data in three dimensions using the variables Breed, Frame, and BkFat. Rotate the coordinate axes in various directions. Check for outliers. Are the breeds well separated in this coordinate system?

(c) Repeat part b using Breed, FtFrBody, and SaleHt. Which three-dimensional display appears to result in the best separation of the three breeds of bulls?

Table 1.10 Data on Bulls

Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
1	2200	51.0	1128	70.9	7	.25	54.8	1720
1	2250	51.9	1108	72.1	7	.25	55.3	1575
1	1625	49.9	1011	71.6	6	.15	53.1	1410
1	4600	53.1	993	68.9	8	.35	56.4	1595
1	2150	51.2	996	68.6	7	.25	55.0	1488
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	1450	51.4	997	73.4	7	.10	55.2	1454
8	1200	49.8	991	70.8	6	.15	54.6	1475
8	1425	50.0	928	70.8	6	.10	53.9	1375
8	1250	50.1	990	71.0	6	.10	54.9	1564
8	1500	51.7	992	70.6	7	.15	55.1	1458

Source: Data courtesy of Mark Ellersieck.

1.27. Table 1.11 presents the 2005 attendance (millions) at the fifteen most visited national parks and their size (acres).

(a) Create a scatter plot and calculate the correlation coefficient.

(b) Identify the park that is unusual. Drop this point and recalculate the correlation coefficient. Comment on the effect of this one point on correlation.

(c) Would the correlation in Part b change if you measure size in square miles instead of acres? Explain.

Table 1.11 Attendance and Size of National Parks

National Park	Size (acres)	Visitors (millions)
Arcadia	47.4	2.05
Bruce Canyon	35.8	1.02
Cuyahoga Valley	32.9	2.53
Everglades	1508.5	1.23
Grand Canyon	1217.4	4.40
Grand Teton	310.0	2.46
Great Smoky	521.8	9.19
Hot Springs	5.6	1.34
Olympic	922.7	3.14
Mount Rainier	235.6	1.17
Rocky Mountain	265.8	2.80
Shenandoah	199.0	1.09
Yellowstone	2219.8	2.84
Yosemite	761.3	3.30
Zion	146.6	2.59

References

1. Becker, R. A., W. S. Cleveland, and A. R. Wilks. "Dynamic Graphics for Data Analysis." *Statistical Science*, **2**, no. 4 (1987), 355–395.
2. Benjamin, Y., and M. Igbaria. "Clustering Categories for Better Prediction of Computer Resources Utilization." *Applied Statistics*, **40**, no. 2 (1991), 295–307.
3. Capon, N., J. Farley, D. Lehman, and J. Hulbert. "Profiles of Product Innovators among Large U. S. Manufacturers." *Management Science*, **38**, no. 2 (1992), 157–169.
4. Chernoff, H. "Using Faces to Represent Points in *K*-Dimensional Space Graphically." *Journal of the American Statistical Association*, **68**, no. 342 (1973), 361–368.
5. Cochran, W. G. *Sampling Techniques* (3rd ed.). New York: John Wiley, 1977.
6. Cochran, W. G., and G. M. Cox. *Experimental Designs* (2nd ed., paperback). New York: John Wiley, 1992.
7. Davis, J. C. "Information Contained in Sediment Size Analysis." *Mathematical Geology*, **2**, no. 2 (1970), 105–112.
8. Dawkins, B. "Multivariate Analysis of National Track Records." *The American Statistician*, **43**, no. 2 (1989), 110–115.
9. Dudoit, S., J. Fridlyand, and T. P. Speed. "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association*, **97**, no. 457 (2002), 77-87.
10. Dunham, R. B., and D. J. Kravetz. "Canonical Correlation Analysis in a Predictive System." *Journal of Experimental Education*, **43**, no. 4 (1975), 35–42.