Example 3.5 (Selecting a random sample) As a preliminary step in designing a permit system for utilizing a wilderness canoe area without overcrowding, a natural-resource manager took a survey of users. The total wilderness area was divided into subregions, and respondents were asked to give information on the regions visited, lengths of stay, and other variables.

The method followed was to select persons randomly (perhaps using a random number table) from all those who entered the wilderness area during a particular week. All persons were equally likely to be in the sample, so the more popular entrances were represented by larger proportions of canoeists.

Here one would expect the sample observations to conform closely to the criterion for a random sample from the population of users or potential users. On the other hand, if one of the samplers had waited at a campsite far in the interior of the area and interviewed only canoeists who reached that spot, successive measurements would not be independent. For instance, lengths of stay in the wilderness area for different canoeists from this group would all tend to be large. ■

Example 3.6 (A nonrandom sample) Because of concerns with future solid-waste disposal, an ongoing study concerns the gross weight of municipal solid waste generated per year in the United States (Environmental Protection Agency). Estimated amounts attributed to x_1 = paper and paperboard waste and x_2 = plastic waste, in millions of tons, are given for selected years in Table 3.1. Should these measurements on $\mathbf{X}' = [X_1, X_2]$ be treated as a random sample of size $n = 7$? No! In fact, except for a slight but fortunate downturn in paper and paperboard waste in 2003, *both* variables are increasing over time.

Table 3.1 Solid Waste

Year	1960	1970	1980	1990	1995	2000	2003
x_1 (paper)	29.2	44.3	55.2	72.7	81.7	87.7	83.1
x_2 (plastics)	.4	2.9	6.8	17.1	18.9	24.7	26.7

■

As we have argued heuristically in Chapter 1, the notion of statistical independence has important implications for measuring distance. Euclidean distance appears appropriate if the components of a vector are independent and have the same variances. Suppose we consider the location of the kth column $\mathbf{Y}_k' = [X_{1k}, X_{2k}, \ldots, X_{nk}]$ of $\mathbf{X}$, regarded as a point in n dimensions. The location of this point is determined by the joint probability distribution $f(\mathbf{y}_k) = f(x_{1k}, x_{2k}, \ldots, x_{nk})$. When the measurements $X_{1k}, X_{2k}, \ldots, X_{nk}$ are a random sample, $f(\mathbf{y}_k) = f(x_{1k}, x_{2k}, \ldots, x_{nk}) = f_k(x_{1k}) f_k(x_{2k}) \cdots f_k(x_{nk})$ and, consequently, each coordinate x_{jk} contributes equally to the location through the identical marginal distributions $f_k(x_{jk})$.