FIGURE 2.1

The discrete probability distribution for all possible rolls of a single fair die. Each roll has the same probability, 1/6.

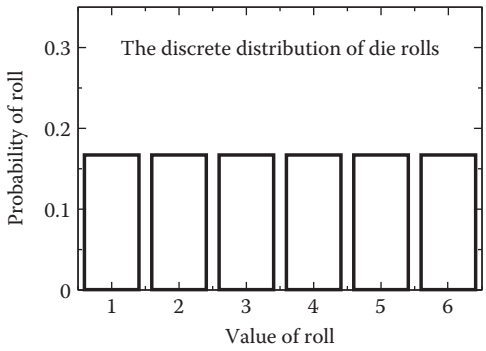

FIGURE 2.2

A Gaussian distribution, also called a normal distribution. We shall see later why Gaussians are indeed "normal" when we discuss the central limit theorem. The pictured distribution has a mean $\langle x \rangle$ of zero and standard deviation $\sigma = 1$. However, you should be able to imagine a distribution with the same shape but different values of the two parameters. The panel (a) shows a standard continuous representation of the Gaussian (i.e., a probability density function), while the panel (b) shows a discretized version. The discrete version can be converted to an approximate density if the probabilities are divided by the constant interval width of 0.04.

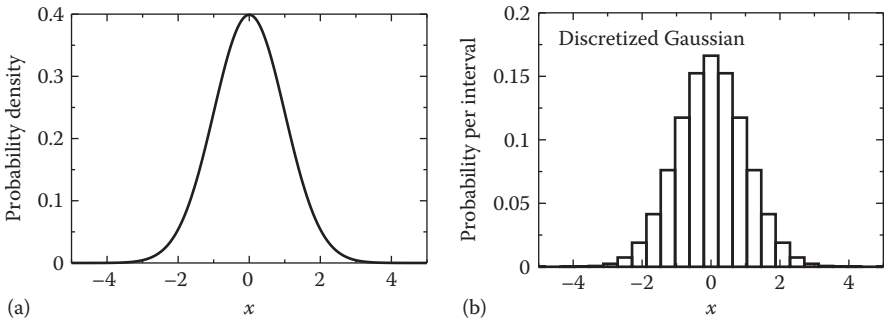

FIGURE 2.3

The distribution of butane's main dihedral. The plotted probability density was calculated from the 10 ns molecular dynamics simulation depicted in Figure 1.3. Although butane is a symmetric molecule, the distribution does not mirror itself about the central value of 180°, due to the finite length of the simulation.

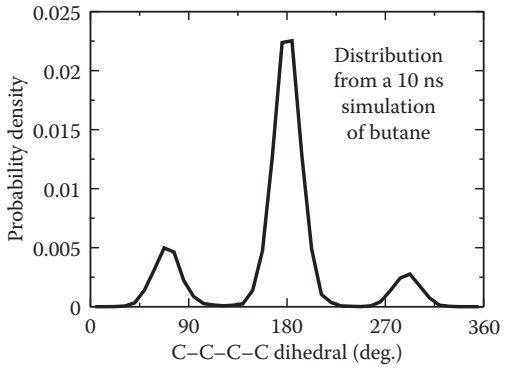

FIGURE 2.4

The cdf of panel (b) is a useful way of visualizing the pdf in (a). The cdf indicates how much probability occurs to the left of its argument in the original pdf. For a smooth distribution with a single peak, the cdf has a characteristic sigmoidal shape. By following particular y-axis values of the cdf down to their corresponding x values, we see that exactly 50% of the probability occurs to the left of zero, and that nearly 90% occurs to the left of 1 ($x < 1$). Correspondingly, nearly 80% of the probability lies between -1 and 1.

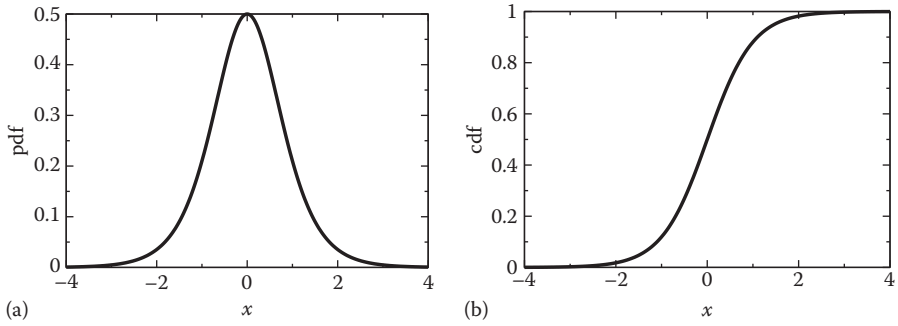

FIGURE 2.5

Histograms produced from several small-sized samples. Panels (a), (c), (e) are based on a rectangular distribution, where $\rho(x) = \text{const.}$ whereas panels (b), (d), (f) result from an exponential distribution given by $\rho(x) \propto \exp(-x)$. Larger samples are progressively included (10, 100, 1000 numbers) from top to bottom. A key point is that it is difficult to determine the type of distribution without a large number of samples. One could even mistake the number of peaks in a distribution from a small sample—for instance, from the top or middle panels.

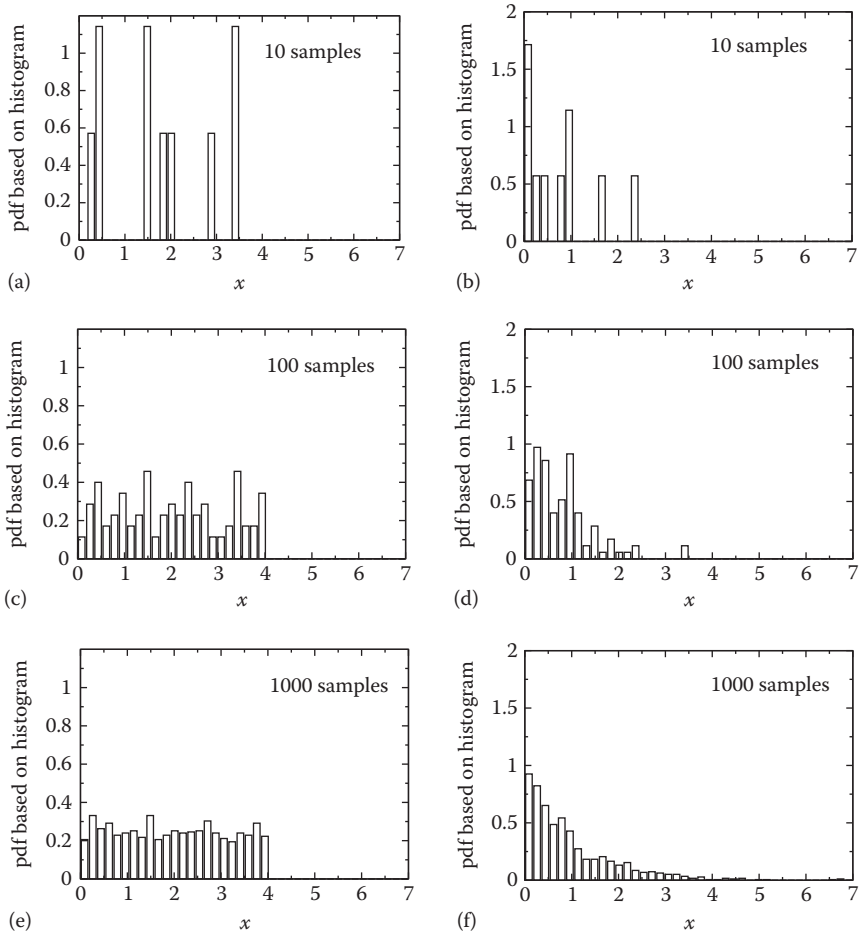

FIGURE 2.6

Correlated sampling in butane. Panel (a) depicts time-correlated or "evolutionary" sampling, which is the typical result of modern computer simulations: every new configuration is not chosen at random from the full distribution shown in panel (b). Rather, new configurations (and hence dihedral angles) evolve from preceding values, and hence the discrete points of the trajectory are appropriately connected with lines. The trajectory at left has not even visited the third state revealed in the pdf based on a much longer simulation (10 ns). In fact, the behavior shown here is typical of any complex system—whether a small molecule or protein. Evolutionary (or dynamical) sampling will only produce the correct distribution after a very long simulation. You can look ahead to Figure 12.4 to see the same effect in a large protein simulation, done on a supercomputer!

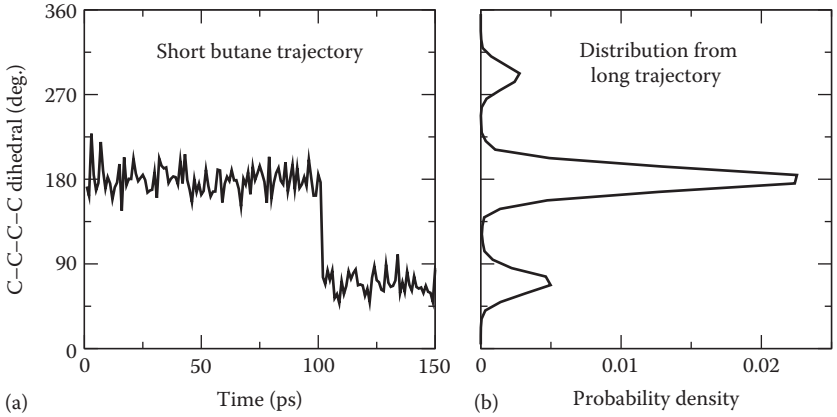

FIGURE 2.7

The natural scale vs. the scale of statistical uncertainty. Every distribution has a natural scale, which is typically quantified by the standard deviation, σ. Here, $\sigma = 3$. This is just the commonsense "width" of the distribution and has nothing to do with how many samples or measurements have been generated from the distribution. By contrast, the statistical uncertainty in a series of measurements from a distribution depends on how many independent samples have been obtained, and is quantified by the standard error (Equation 2.22). With a large number of measurements (independent samples), the statistical uncertainty can become much smaller than the standard deviation.

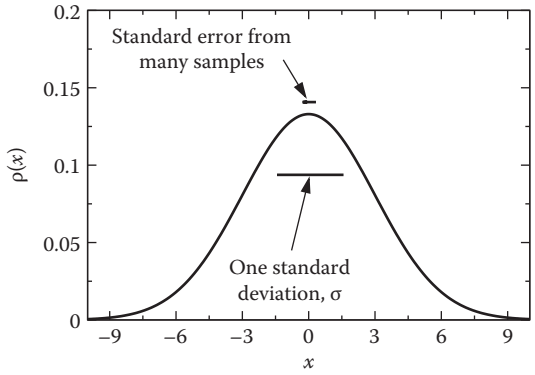

FIGURE 2.8

An asymmetric distribution. Specifically, a gamma distribution is shown. In contrast to the Gaussian distribution (Figure 2.2), values to the left and right of the mean are not equally probable.

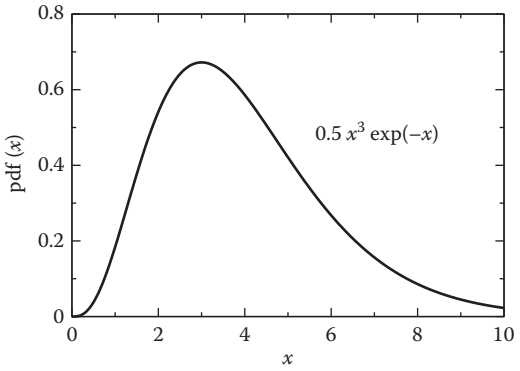

FIGURE 2.9

The projection procedure and its pitfalls. A two-dimensional probability distribution $\rho(x, y)$ is projected separately onto both x and y axes. The original ρ may be thought of as representing piles of leaves on the ground whose height is not constant but "tails off" at the edges. Evidently, the projection onto y correctly captures the three distinct piles, though the projection onto x does not. For a complex system with many variables, such as a biomolecule, it may be difficult or impossible to determine a low-dimensional set of coordinates onto which an informative projection can be made.

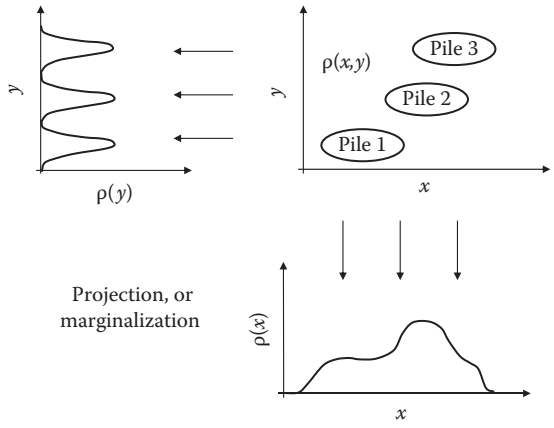

FIGURE 2.10

Examples of linearly correlated data (a), nonlinearly correlated data (b), and uncorrelated data (c). Note that the data set in (b), which is clearly correlated, would yield a Pearson coefficient of zero by Equation 2.29.

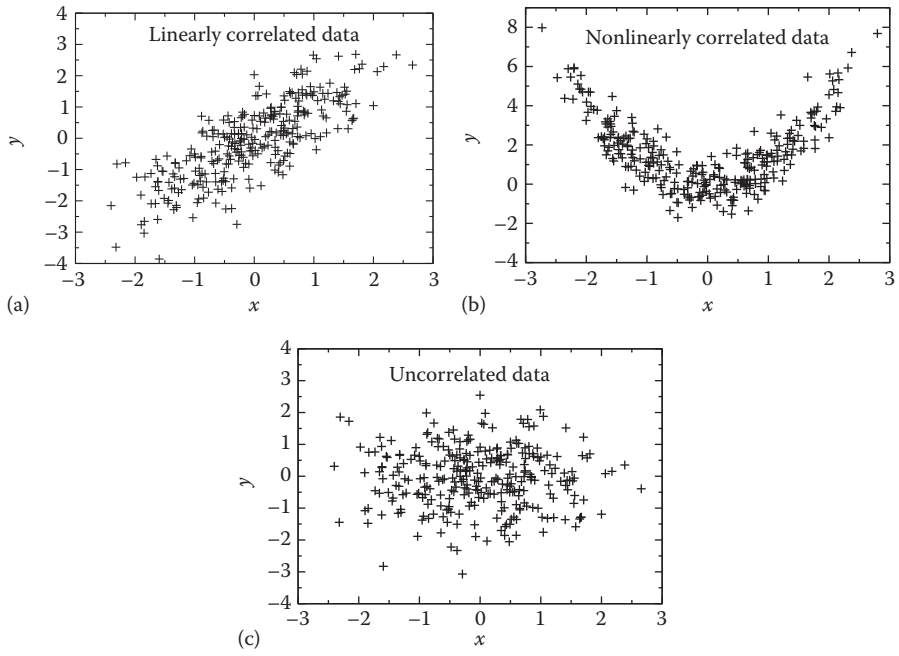

FIGURE 2.11

Molecular correlations, weak and strong. Scatter plots are shown in which each point represents a pair of coordinate values based on a single configuration from a long computer simulation. (a) For butane, subtle correlations are visible in two regards: the mean C–C–C bond angle of the central *trans* state is slightly lower than that of either of the *gauche* states; also, the *gauche* distributions are slightly diagonal. Correlations in alanine dipeptide (b) are stronger. This can be seen by comparing ψ distributions at different fixed ϕ values.

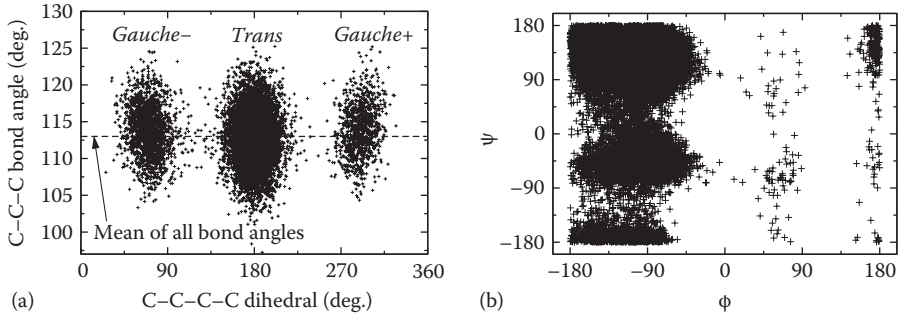

FIGURE 2.12

Constructing a conditional probability. Given the condition that $1 < x < 2$, which means only the points between the dashed vertical lines are considered, one could construct a corresponding distribution for y by Equation 2.32. In this case, without normalization, we have $\rho(y|1 < x < 2) \propto \int_1^2 dx\, \rho(x, y)$.