2.4.3 Statistical Independence

We will first consider the absence of correlation, which is often called statistical independence (or simply "independence") and is quite important in itself. An example is that the number of pennies in the pocket of a professor in Pittsburgh should be statistically independent of the price of gold in London on any given day. In mathematical terms, the distribution of a set of statistically independent variables factorizes in a simple way, namely, if x and y are independent, then

$$\rho(x, y) = \rho_x(x)\, \rho_y(y). \tag{2.26}$$

Here I have included subscripts on the single-variable distributions to emphasize that they are different functions.

PROBLEM 2.32

Show whether or not the two-dimensional Gaussian of Problem 2.30 can be written in the form of the right-hand side of Equation 2.26.

PROBLEM 2.33

(a) Given a distribution $\rho(x, y)$ as in Equation 2.26, write down the integrals defining ρ_x and ρ_y. (b) Note the connection to projection, but explain why a projected distribution might or might not describe an independent variable.

In general, you should assume variables are *not* independent unless you have a good reason to believe they are!

2.4.4 Linear Correlation

While there are many subtle types of correlation, as will be sketched below, the simplest linear correlations are easy to understand and quantify. A typical example that commonly appears in the literature is the comparison between results for the same quantities obtained by different means. For instance, you may numerically estimate a set of quantities $A(j)$ for $j = 1, 2, \ldots$ (denoted $A_{\text{calc}}(j)$) that also can be measured experimentally—$A_{\text{meas}}(j)$. Each j value could represent a different ligand molecule and A could be its binding affinity to a given protein. As sketched in Figure 2.10a, for every j one can plot a point whose location is defined by the coordinates $(x, y) = (A_{\text{calc}}(j), A_{\text{meas}}(j))$. This is called a "scatter plot." Perfect agreement would mean exact equality between the measured and calculated values, in which case all points would fall on a diagonal line passing through the origin. It is clear that one can get a visual feel for the level of agreement or correlation using a scatter plot.

A "correlation coefficient" quantifies the degree of linear correlation. (Let us be clear that linear correlation is not the whole story, as depicted in Figure 2.10.) The mathematical basis for the correlation coefficient can be understood as a contrast to the case of statistical independence, given above in Equation 2.26. In particular,

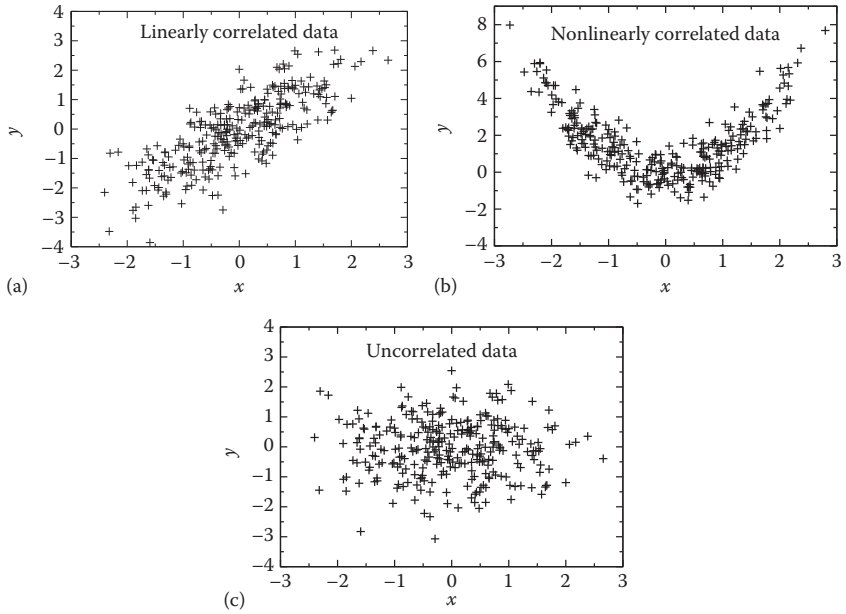

FIGURE 2.10 Examples of linearly correlated data (a), nonlinearly correlated data (b), and uncorrelated data (c). Note that the data set in (b), which is clearly correlated, would yield a Pearson coefficient of zero by Equation 2.29.

imagine computing the average of the product of two independent or uncorrelated variables. We would be able to factorize this product as follows:

$$\text{Uncorrelated: } \langle x\,y\rangle = \int dx\,dy\,x\,y\,\rho(x,y) = \int dx\,x\,\rho_x(x)\int dy\,y\,\rho_y(y)$$
$$= \langle x\rangle\,\langle y\rangle\,. \tag{2.27}$$

The factorizability for independent variables seen in Equation 2.27 suggests a simple route for quantifying correlation. In particular, if the equality in Equation 2.27 does not hold, then the two variables are correlated. Noting the identity $\langle xy\rangle - \langle x\rangle\,\langle y\rangle = \langle (x - \langle x\rangle)\,(y - \langle y\rangle)\rangle$, we can define linear correlation as occurring whenever

$$\langle (x - \langle x\rangle)\,(y - \langle y\rangle)\rangle \neq 0\,. \tag{2.28}$$

Note that the deviations from the respective means of x and y arise naturally here.

Yet as they are written, Equations 2.26 and 2.28 are useless for practical purposes. There is always statistical and other error in our calculations (based on finite data), so one could never convincingly demonstrate statistical independence using these relations: What would a "small" deviation from zero mean? In fact, what is small anyway?!?