The histogram is a nonparametric tool. It estimates a generic \( p(y) \) without making any assumption about any named form (normal, Poisson, etc.) of the distribution. Sometimes it is handy to compare the histogram estimate of \( p(y) \) with a distribution having a particular named form, to see whether such a named model would be reasonable. If the histogram and the named distribution are very similar in appearance, then you can be more confident in using the named distribution as a model.

As you can see from comparing the previous histograms with \( n = 6 \) versus \( n = 1000 \) observations, the histograms look a lot better with larger \( n \). This is no surprise, since probability estimates are better with larger \( n \)—just think of flipping a bent coin \( n = 6 \) times versus \( n = 1000 \) times. So, larger \( n \) is better, but how large is “large enough?” Here is another ugly rule of thumb:

**Ugly Rule of Thumb 4.2**
With a sample size of \( n \geq 30 \), the histogram is an adequate estimate of the distribution \( p(y) \) that produced your data. Larger \( n \) provides better accuracy.

**Interpreting Histograms:**
1. Notice the range (on the horizontal axis) of the observed values.
2. Notice the approximate center of the distribution of the observed values.
3. The histogram is an estimate of the pdf. Look for indications of symmetry or asymmetry or any special unusual features such as bimodality (i.e., two distinct peaks).
4. Often the histogram is used to assess the adequacy of the normal model. If so, look for rough symmetry and bell shape of histogram. Do not look for perfection though: Even data* produced by a normal distribution will not have a perfectly bell-shaped histogram due to randomness.
5. Discuss the sample size. With larger sample sizes, histograms become better approximations to the distribution that produced the data.

If you use the histogram—or the quantile–quantile plot discussed later in this chapter—to assess a particular distribution form, like normality, you should never conclude “the data are normally distributed” or “the data are not normally distributed.” Both statements are meaningless, as data can never be normally distributed, or have any other continuous distribution, for that matter. They are just discrete values, even if measured to many decimals. So any claim that the data are or are not normally distributed is purely nonsense.

The question of normality, or of any other distribution form such as the exponential, Poisson, etc., is a question about the process that produced your data. (Model produces data.) It is not a question about your data set. The data do shed light on your question, “Which distribution produced my data?” but they do not answer it. Data reduce the uncertainty about the unknown parameters. They do not eliminate your uncertainty.

**Example 4.4 Estimating the Distribution of Stock Market Returns via the Histogram**
In Chapter 1, we introduced the concept of the financial return, which is the relative price change from one day to the next. We showed in Example 1.7 how you can create potential future stock price trajectories when you know the distribution of the returns. Example 1.7 assumed a normal distribution. Is that a reasonable assumption? One way to check whether you can assume a normal distribution to produce returns is to examine historical return data, draw the histogram, and superimpose (via software) a normal distribution. Data from the Dow Jones Industrial Average (DJI/A) are freely